

SECTION II.—GENERAL METEOROLOGY.

ELEMENTARY NOTES ON LEAST SQUARES, THE THEORY OF STATISTICS AND CORRELATION, FOR METEOROLOGY AND AGRICULTURE.

By CHARLES FREDERICK MARVIN.

CONTENTS.

	Page.
Introduction.....	551
Explanation of terms:	
Average, mean, normal.....	552
Errors, residuals, departures.....	552
Correction.....	552
Probability curve.....	552
Frequency curve.....	552
Distribution, frequency distribution.....	552
Least squares.....	553
Probable error; probable variation.....	553
Elements of the laws of probability.....	553
Theory of errors and method of least squares.....	553
Normal frequency curve.....	554
Properties of the probability curve.....	555
Areas of normal error curves.....	556
Probable error.....	556
Standard deviation.....	557
Evaluation of "h".....	557
Probability integral.....	558
Most probable departure; mode.....	558
Median.....	558
Procedure in statistical studies.....	558
Moment of curves.....	558
Mathematical curve fitting.....	558
Correlation.....	560
Value of the correlation coefficient.....	560
Examples.....	562
Practical calculation of the standard deviation.....	563
Probability of a given departure.....	566
Summary and conclusion.....	567
Literature on correlation.....	568

Much attention is now being given by a few to the application of the laws of probabilities and of the theory of statistics, least squares, and correlation to studies in meteorology, especially in its relations to agriculture. These mathematical agencies may be made very useful to meteorologists and others in the analysis and comparison of the data of their respective sciences and in the determination of possible relations between the several phenomena or quantities they discuss.

The customary meteorological and climatic tables tell us a great deal about the mean temperature and average weather conditions of this or that locality and for the several days, months, seasons, etc. The departures from the normal, the extremes, and ranges of the conditions are also fully specified. Nevertheless, many realize how these data often fail to bring out climatic characteristics that impress themselves strongly upon our physiological or psychological sensations. A possible explanation of some of these anomalies may be found in recognizing that the *average* temperature, for example, or the *average* conditions of any kind are not necessarily the most frequent or the most probable conditions. The average or the mean *may* be the most frequent in *some* localities or at certain times and seasons, whereas it is entirely possible and in fact quite probable that in another locality or on another occasion the most frequent temperature may be either above or below the mean or normal.

The *standard deviation* is another mathematical concept that may be employed to express a characteristic or peculiarity of a given climate and likewise claims attention. Different localities and different elements may show marked differences in these characteristics and such

differences can not fail to be accompanied by psychic and physiological effects on man and by corresponding effects on other organisms. All these climatic features and peculiarities can be fully disclosed only by a proper application of the laws of statistics to the several problems in hand. Vast quantities of suitable data are already available and await analysis and discussion of the kind indicated.

Many interesting questions may be answered with assurance by means of the mathematical methods referred to; for example:

1. In the long run, how many days in December at Washington, for example, will the minimum temperature fall below freezing?

[Some will be content to count up the times for a series of years. A far better result may be obtained, however, by a proper application of the theories of statistics.]

2. What will be the most frequent temperature at a given locality and for a specified interval of time?

3. Below what temperature will the minimum for the day at a designated place and period of time be just as likely to fall as not?

4. How many times in a month will the temperature be the same as the average?

5. What percentage of maximum temperatures for a given month and locality will be 10°, 15°, or 20° above the normal?

6. How do different localities differ in one or more of the particulars indicated in the foregoing questions?

7. Are differences in the particulars cited reflected in the growth of crops or the welfare and comfort of individuals, and, if so, what are the economic and hygienic aspects thereof?

Weather conditions rainfall, river stages, crop yields, and many other important phenomena of meteorology and agriculture can be analyzed and set forth with advantage by the mathematical methods indicated.

Prof. Karl Pearson, of University College, London, has done much to systematize and reduce to a practical working basis the intricate mathematical processes that must be employed in order to reach exact and rational results in the discussion of large groups of statistical data. His important contributions to this science should be consulted by every careful student.

Except for the safe and certain basis afforded by the science of statistics for the risks of insurance, for example, this great industry in all its manifold applications of the present day would be either a costly investment to the policyholder, or a losing venture to the company, and not the perfectly safe and equitable business enterprise it now is.

Biologists, anthropologists, and other students of statistics generally now employ the exact mathematical methods of this comparatively new science with great advantage in their several fields of work. A few meteorologists must also be mentioned in this list, but as yet little has been done, especially in American climatology and agriculture,¹ to use this science to the fullest possible extent in solving many of the practical problems of applied meteorology and scientific agriculture.

¹ Prof. W. J. Spillman has recently made important practical applications of statistical theories to the problems of farm management as dependent on weather factors. His assistants, Messrs. W. G. Reed and H. R. Tolley, are also making investigations of climatic and agricultural statistics from the mathematical standpoint, and useful results may be expected. See MONTHLY WEATHER REVIEW, April, 1916, 44: 197; June, 1916, 44: 354.

Many who are interested or engaged in the study of the data of weather, climate, and crops are but slightly acquainted with advanced mathematics. Such are frequently dismayed by the formidable array of mathematical symbols commonly employed to demonstrate the theories of errors or variations and fail to follow and learn the methods and processes by which these theories may be applied in the solution of many practical problems. Hence, it seems a useful purpose will be served by stating in simple verbal terms, as far as possible, some of the generalizations warranted by the more elegant mathematical demonstrations and by illustrative examples of the processes of computation indicated by the equations of this branch of science.

With this object in view it is proposed to discuss briefly several of the more or less technical terms now in use and illustrate by simple examples certain important methods of computing results.

EXPLANATION OF TERMS.

Average, mean, normal.—Mathematically, the quantities to which these names are applied in dealing with statistical data are essentially the same, namely, the quotient found by dividing the sum of a series of values by the number of values. In ordinary usage there is no essential difference in the significance of average and mean, although the latter sometimes means the value midway between two extremes, whereas the average has only the mathematical significance of the sum of the observations divided by the number. The word *normal*, however, as used in meteorology, is supposed to have a special significance, which can not be more precisely defined than to say that it is the average value of a long series of observations. Just how long the series must be to make the average a real normal can not be known, except possibly to say that the average becomes the normal when its value ceases to change appreciably with increased length of record. This is altogether a question of the importance or significance of small changes in the value of the average. Notwithstanding its vagueness of meaning the word *normal* is a convenient one to distinguish the average of a few results from the mean of a considerable number, or even the greatest number of values available.²

Errors, residuals, departures.—These terms often have a similar significance, but in other cases important differences of meaning must be recognized. All measurements and observations of physical magnitudes, as also the observation and determination of the elemental terms constituting statistical data, show variations when more than one observation or determination are made. In general it is impossible to learn the true magnitude. When used in its strictest sense, the word *error* means the difference between some particular observed value and the true value of a datum. Since the true value, and therefore the true error, can perhaps never be known, it is customary to adopt in place of the true value some closely approximate, or best, or most probable value. Careful writers will then use the word *residual* to mean the difference between some particular observed value and the most probable value. The word *error*, however, is often loosely used to have the same significance as *residual*.³

Errors and residuals may be considered as variations of repeated observations or measurements, from the true

or the approximate value of a given magnitude *which in effect is assumed to remain fixed and invariable*. While the determination of each element of statistical data—*c. g.*, the velocity of the wind, the temperature of the air, the amount of rainfall, etc.—is, strictly, subject to errors of observation, nevertheless such data are also subject to an additional cause of variation due to the progressive and actual change in the true value of the wind velocity, the temperature of the air, the rainfall, etc. In such cases the *errors* of determination can not as a rule be eliminated or separately analyzed, but they are generally small and insignificant in comparison with the larger variations of the elemental datum itself.

The term *departure* may be reserved to designate the difference between the mean value and some one of the values of a quantity *which is more or less continuously changing its magnitude*. It is not at all necessary, however, to form the departures with respect to the mean. Indeed, it is generally easier and better to use some arbitrary value, as the computations can thus be carried out with greater facility and even more exactly. (See p. 563.)

Errors and residuals represent inexactness and imperfection in determining quantities of a fixed magnitude. *Departures* show the nature and amount of variations in a quantity of changing magnitude.

Correction.—The numerical value of a correction in a given case is always exactly the same as that of an error or residual, or departure, but it has the opposite algebraic sign. It is the quantity that must be added algebraically to an *observed value* to deduce thereby the *true* or *most probable* value. The word has no particular use in the study of statistical data, but is explained here to bring out the difference between *corrections* which are very frequently employed in adjusting observations, measurements, and data generally and the *variations* or *departures* which constitute the ground work of statistics.

Deviation is a word which is also used to refer to the amount of departure from a mean or arbitrary base value.

Standard deviation is the name applied to a particular value of the departures in a group of data and is extensively employed in statistics as an index of the variability of the group and for other purposes. Its importance or usefulness depends chiefly upon its mathematical significance, as will be brought out more fully in what follows.

Probability curve.—A curve of the general character shown in figure 1 and representing the law of frequency with which errors or residuals of different magnitude occur. The height of any point on the curve above the base line is proportional to the number of errors or residuals which have the value represented by the distance of the point from the vertical axis of symmetry. *Normal distribution, normal curve of errors* are other names applied to the probability curve.

Frequency curve.—A curve exhibiting the law of frequency of occurrence of departures of various magnitudes from a base or reference number. The curve may be identical in character with the probability curve, but unsymmetrical curves of various kinds may also be required to exhibit particular kinds of data, as will be indicated later.

Distribution, frequency distribution.—Expressions referring to the grouping of departures as exhibited by a frequency curve or diagram. An easy and expeditious manner of charting a frequency distribution is the dot system shown in figures 15 and 16. When the departures have been calculated the dot chart can be rapidly built up and the result is graphic and effective. In fact, the

² On this subject of "mean" and "average" see also this REVIEW, Jan., 1915, 43: 24; Aug., 1895, 23: 294.—C. A. Jr.

³ *Merriman, M.* Method of least squares. New York, 1915. p. 5.

chart is often just as easily formed directly from the observations, thus reducing to a minimum the labor of computing and tabulating numerical departures.

Least squares, sum of squares.—These expressions refer to the sum of the squares of the errors, or residuals, or departures for a given group of data. The magnitude of the sum will depend on the base number with reference to which the departures are taken. Some one base number will cause the sum of the squares to be a minimum, hence the expression "least squares." When the data represent physical measurements, that value which makes the sum of squares the least is the most probable value and therefore the most accurate value obtainable from the data discussed. The arithmetical mean is the most probable value in the case of observations of equal accuracy.

In dealing with statistical data the most probable value does not have the significance of the most accurate value, but simply is the most frequent value. If the "distribution" is unsymmetrical the arithmetical mean will then no longer be the most frequent value, but some other value will be the most frequent. The name *mode* is now used to designate the most frequent value.

Probable error, probable variation.—These expressions are mathematically identical, at least when the "distribution" is normal, nevertheless there is clearly an appropriate distinction between them. The probable error has reference to a group of measurements of a fixed quantity and designates that error than which half of the errors are greater and the remaining half are less. The probable variation refers to the changing values of a variable quantity and is the name of that magnitude of the departure which has the middle value, that is, half of the deviations in a given group of data are greater, and half are less, than the probable variation.

Variant.—This term is frequently used to designate in a general way any group of data exhibiting variation. In meteorology, for example, the rainfall for a given period and locality might be referred to as the *variant*, and a given single value as a *variant*.

ELEMENTS OF THE LAWS OF PROBABILITY.

Let m = number of ways in which a certain simple event can happen,

n = number of ways in which the event can fail, then

$m + n$ = total number of cases that can occur.

If the probability that the event will happen be called p , then

$$p = \frac{m}{m+n}.$$

If $n = 0$, $p = \frac{m}{m} = 1$, that is, it is certain the event will happen since 1 is the measure of certainty.

If $m = 0$, $p = \frac{0}{n} = 0$; that is, it is certain the event will fail, zero being the measure of complete failure.

If a compound event occurs when two or more simple events happen together, then the probability of the compound event is the product of the probabilities of the simple events.

THE THEORY OF ERRORS AND THE METHODS OF LEAST SQUARES.

These expressions designate a group of theories and mathematical methods based on the laws of probabilities, all originally developed chiefly by astronomers, mathematicians, and geodesists for the purpose of provid-

ing rational methods of comparing, adjusting, combining, and harmonizing many different measurements or observations of one and the same quantity so as finally to ascertain the best, or the most accurate, that is, the most probable value that could be deduced from all the observations.

This theory, however, has a far wider application than simply to errors of measurement. In meteorology and climatology it may be a valuable mathematical agency for indicating the nature and amount of variation in the data commonly discussed.

The theory of errors, in fact, deals with only one of the special and relatively simple cases that fall within the domain of the more general theories embracing all classes of statistical data subject to variations. The theory of statistics is the general science. The theory of errors is the special case. Care must be exercised, therefore, that formulae and methods appropriate for the analysis and discussion of errors of measurements are not applied to data whose laws of frequency and variation may be quite different from those upon which the theory of errors is constructed.

It is also well known that the theory of errors—or "of departures from a mean" as the theory might be called for meteorologists—applies strictly only to a very large number of observations or values. The smaller the number of cases the less reliable are the deductions by this theory. As already stated, the principles of the theory were formulated with respect to errors of measurements and observations; nevertheless, under proper restriction they may also be applied to many studies and discussions of statistical data generally. In these latter cases the question at issue is not one of errors of observation, in the proper sense of that expression. Meteorological observations and statistical data generally are all subject to errors of determination or measurement, and it is often impossible to adequately eliminate these causes of variations from the results under discussion. Nevertheless the pure errors of observation will often be small and in the application of the theories now under consideration the "residuals" will be departures from a mean value or arbitrary base number and not errors of determination. The departures represent chiefly the variations in the values of a given datum with respect to some average or mean value that it may be convenient to employ as an arbitrary reference value, but includes nevertheless the actual errors of observation.

Take, for example, the mean temperature of January, or any other month for a long series of years, or, better the mean temperature for any day of the year. This datum tends to be a constant; nevertheless, values from year to year differ more or less from the mean, sometimes higher, sometimes lower. Some months or days and some localities show greater variations than other months and other localities. This variation is a more or less fixed characteristic of the given month or of the locality and some standard method of generally measuring and designating such variations is needed. The method of least squares supplies such a standard method or unit of measurement of variation. This unit is known as the probable error or, giving it a more appropriate name, the probable variation. The term standard deviation is also a name for another measure of this variation that serves to define in exact measure the variation of similar climatic, meteorological, or statistical data to which the methods of least squares may be applied. These terms will be described and defined more fully hereafter.

At the very outset of any general attempt to apply the methods of least squares to some particular data other

than to questions involving only errors of observation, it is highly important to make sure that the theory can properly be applied to the particular data in question and in the manner proposed. Whether this application can be made or not depends on whether the law of the frequency of occurrence of the large and the small deviations from the mean in the data is the same general law as for errors of observations. That is, the extent to which this application is justified depends on how closely the so-called frequency curve for the data in question corresponds to the frequency curve for errors of observations. We must therefore recognize just what are the characteristics and limitations of the normal curve.

Normal frequency curve.

The essential features of the law of frequency of occurrence of errors is represented by what is often called the "normal frequency curve." This law of frequency is fixed by three principal conditions which must be satisfied, namely, (1) very small errors occur with the greatest frequency; (2) very large errors rarely occur; (3) positive errors (values of the data in excess of the true value) and negative errors (values less than the true value) are equally numerous. This latter condition makes the frequency curve symmetrical about the central vertical axis.

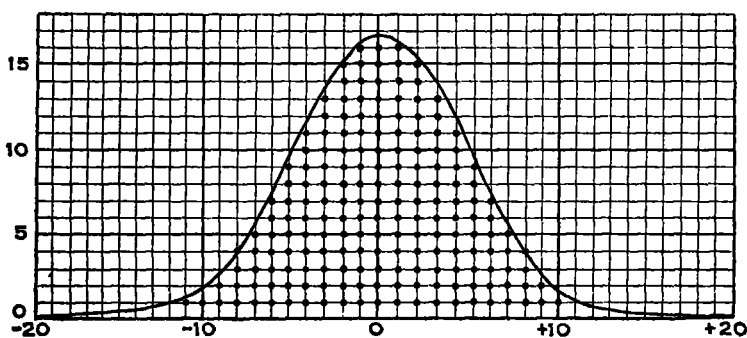


FIG. 1. Normal frequency curve (conventional). Dots represent observations or values assumed to be very numerous and crowded close together.

Figure 1 represents a normal frequency curve having these characteristics. Each dot may be considered to represent one observation or one value of the departure from the mean, the amount of that departure being shown by the distance of the dot from the central line. All the dots comprise all the observations.

Scientists have determined that the normal probability curve of the kind shown in figure 1 is best represented by a certain exponential equation which written in a form commonly found in the textbooks⁴ is as follows:

$$y = \frac{h}{\sqrt{\pi}} e^{-hx^2} \quad (1)$$

in which y represents the probability of a given value or deviation from the mean represented by x . The quantities π and e are well-known mathematical constants, namely, $\pi=3.1416$, the ratio of the circumference to the diameter of circles and $e=2.71828$, the base of the Napierian system of logarithms. These quantities are the same for all kinds of data whatsoever. The one remaining factor h is a constant depending on the particular data under discussion. This quantity h is sometimes called "the measure of precision."

When h is small the curve is low and spreads out laterally as a or b in figure 7. For such a case the observations are inexact or the data vary greatly, and relatively large departures from the mean occur frequently. Such data lack precision and exhibit wide variations. A large

value of h , however, signifies that the measurements or data are grouped closely about the mean value, small deviations are numerous and large ones seldom occur. The curve in this case will have such a form as shown at c or d in figure 7.

According to equation (1) the frequency curve has branches on the right and left extending to infinity, but errors or departures even approaching infinite magnitude do not occur. Hence we must recognize that equation (1) only approximately represents the real case in nature. Nevertheless, the mathematical curve can be made to lie so close to the base line as to practically coincide with it at the limits of the range of data.

It was mentioned with respect to figure 1 that all the dots comprise all the observations. In an analogous but more generalized sense the entire area under the curve including its branches to infinity, represents all possible errors; and the area under the curve between any two defining vertical lines in proportion to the whole area represents the probability of the occurrence of errors of sizes between the defining lines.

When we undertake to study the variation of statistical data generally, it is quite obvious we must expect to find frequency curves widely different from the normal error curve. Pearson and others have classified many of these curves and determined their general equations. A few extreme types are selected here to illustrate the possibilities and are shown in figures 2-8.

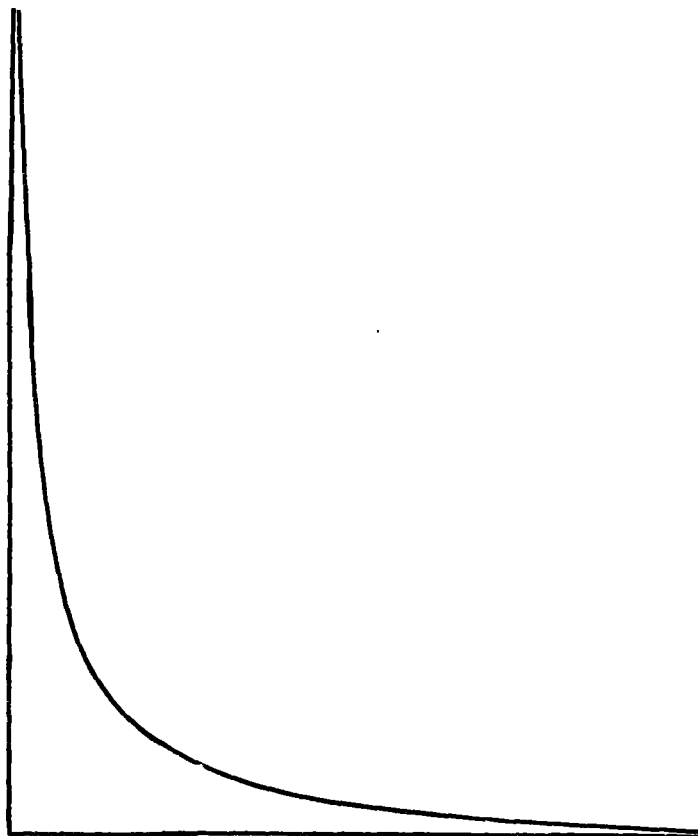


FIG. 2. Extreme asymmetrical frequency curve. Illustrates frequency of deaths from infantile diseases.

Figure 2 shows the extreme asymmetrical⁵ type illustrating, for example, the frequency of deaths at different

⁴ Unfortunately the designation "asymmetrical" is frequently applied by high authorities to curves which are but slightly unsymmetrical. Clearness of expression and consistency seem to justify restricting the use of "asymmetrical" to cases of total absence of symmetry, analogously with the application of "aperiodic" to motions of harmonic character but devoid of period. The word "unsymmetrical" may then be employed to designate frequency curves that are neither perfectly symmetrical, as in the case of the normal curve of errors, nor yet completely devoid of symmetry as in the case of the truly asymmetrical curve.

⁵ Comstock, George C. Methods of least squares. Boston, 1889, p. 5.

ages for certain infantile and children's diseases, like diphtheria, etc. A curve of this type shows also the distribution of wealth.

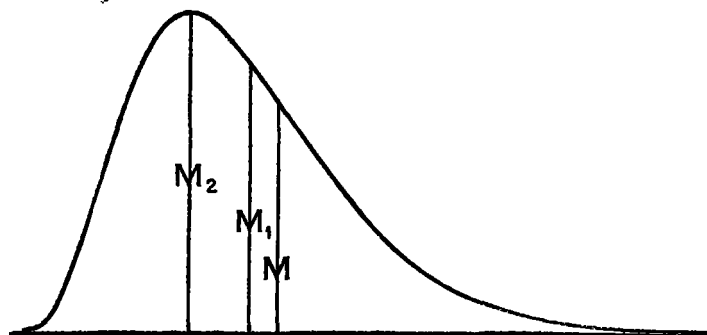


FIG. 3. Unsymmetrical frequency curve.
 M_1 —The mean or average value.
 M_2 —The median or value of middle magnitude.
 M_3 —The Mode, the most frequent value.

Figure 3 shows an unsymmetrical⁵ curve representing many classes of data, exhibiting varying degrees of asymmetry. Such curves may be classified into several distinct types according to significant mathematical criteria distinguishing between their equations.

Skewness is another term employed to indicate lack of symmetry in a frequency distribution.

Some kinds of data are most conveniently analyzed in classes or subgroups. For example, the different amounts of rainfall at a station may be shown in half-inch groups. In such cases a frequency diagram like figure 4 is called a *frequency polygon*. In these cases the frequencies are imagined to be concentrated on the central ordinates of the successive groups, and this method is frequently spoken of as the method of loaded ordinates. Another illustration of a diagram of this character is found in figure 18. The same data may be represented also by a system of rectangles like figure 5, which is sometimes called a histogram.

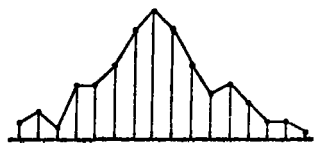


FIG. 4. A frequency polygon exhibiting the irregularities of actual data.

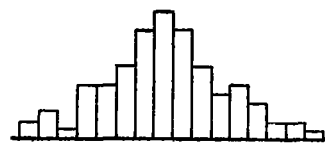


FIG. 5. A frequency diagram outlined by rectangles sometimes called a histogram.

Figure 6 is an unusual type of frequency curve drawn from climatic data and showing the frequency of estimated cloudiness at Breslau⁶ during the 10 years 1876-1885. A completely clouded sky is the most frequent condition, while a perfectly clear sky is the condition next in order of frequency. Intermediate percentages of cloudiness are more rare.

Properties of the probability curve.—The normal error curve has a number of interesting properties, some of which will be briefly considered.

The fundamental equation of the curve is

$$y = \frac{h}{\sqrt{\pi}} e^{-hx^2} \quad (1)$$

Since π and e have the same values for all kinds of data, and since we may, if we desire, use the same scale for x in representing different classes of data, it results that a family of curves such as shown in figure 7 suffices to represent every possible group of data that can be represented by the normal frequency distribution. Each curve represents a given group of data with a corresponding value of h .

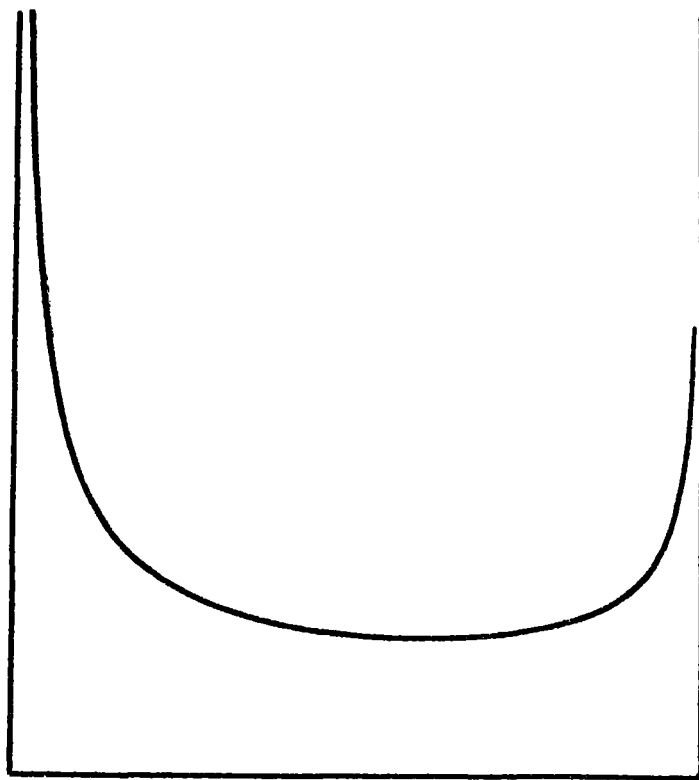


FIG. 6. Unusual frequency-curve representing cloudiness.

The curves in figure 7 were drawn from the data in Table 1, below, giving values of y for a series of values of x in equation (1) and various values of h . The logarithmic form of equation (1) is employed for easy computations, thus:

$$y = \frac{h}{\sqrt{\pi}} e^{-hx^2} \quad \text{or} \quad \log y = \log \frac{h}{\sqrt{\pi}} - hx^2 \log e.$$

Putting $\log \frac{h}{\sqrt{\pi}} = a$ and $h^2 \log e = b$, we get the simple equation

$$\log y = a - bx^2 \quad (1a)$$

If a calculating machine of the "Brunsviga" type is available, values of the second member of the equation may be rapidly computed by first setting up a in the product roll, then b in the number roll; finally the successive subtraction of bx^2 from a by properly operating the machine for successive values of x , gives values of $\log y$ directly in one operation.

⁵ Yule, G. Udny. An introduction to the theory of statistics. London, 1912, p. 103.

TABLE 1.—Values of a , b , and y , for values of x and different values of h , computed by equation (1a).

h	a	b	x											
			0	2	4	5	6	8	10	15	20	25	30	
0.05	8.45040	0.00108	y	y	y	y	y	y	y	y	y	y	y	
0.10	8.75142	0.00434	0.0282	0.0279	0.0271	0.0265	0.0258	0.0240	0.0220	0.0181	0.0104	0.0059	0.0030	
0.15	8.92752	0.00677	0.0564	0.0542	0.0481	0.0439	0.0391	0.0298	0.0208	0.0059	0.0010			
0.20	9.05246	0.01737	0.0846	0.0773	0.0590	0.0482	0.0375	0.0200	0.0089	0.0005				
0.25	9.14937	0.02714	0.1128	0.0862	0.0595	0.0415	0.0267	0.0087	0.0021					
0.30	9.23855	0.03909	0.1410	0.1038	0.0518	0.0296	0.0149	0.0029	0.0003					
0.35	9.29590	0.05320	0.1693	0.1181	0.0401	0.0178	0.0066	0.0005	0.0000					
0.40	9.35349	0.06949	0.1975	0.1210	0.0278	0.0022	0.0021							
0.45	9.40464	0.08791	0.2257	0.1190	0.0174	0.0041	0.0007							
0.50	9.45040	0.10857	0.2539	0.1129	0.0099	0.0016	0.0002							
			0.2821	0.1038	0.0052	0.0005								

Area of normal error curves.—It was pointed out in connection with figure 1 that the dots comprise all the observations, and as these frequency curves are always assumed to apply exactly only to very large numbers of observations, it is customary to think of the dots as being extremely numerous and crowded very closely together, so that, in fact, the *area* under any particular curve, including its branches to $+$ and $-$ infinity, represents every possible error or departure from the mean that can occur. Since every possible value of the departures is *certain* to occur in the long run, and since 1 is the symbol of certainty, it must follow that the area under any one of the error curves should be exactly unity; also that all the curves should have the same area. This may be proved to be the case by the methods of the calculus, but the fact may also be approximately verified by counting the small rectangles under any of the curves in figure 7. On the scale of that diagram the area of one of the little rectangles has the value

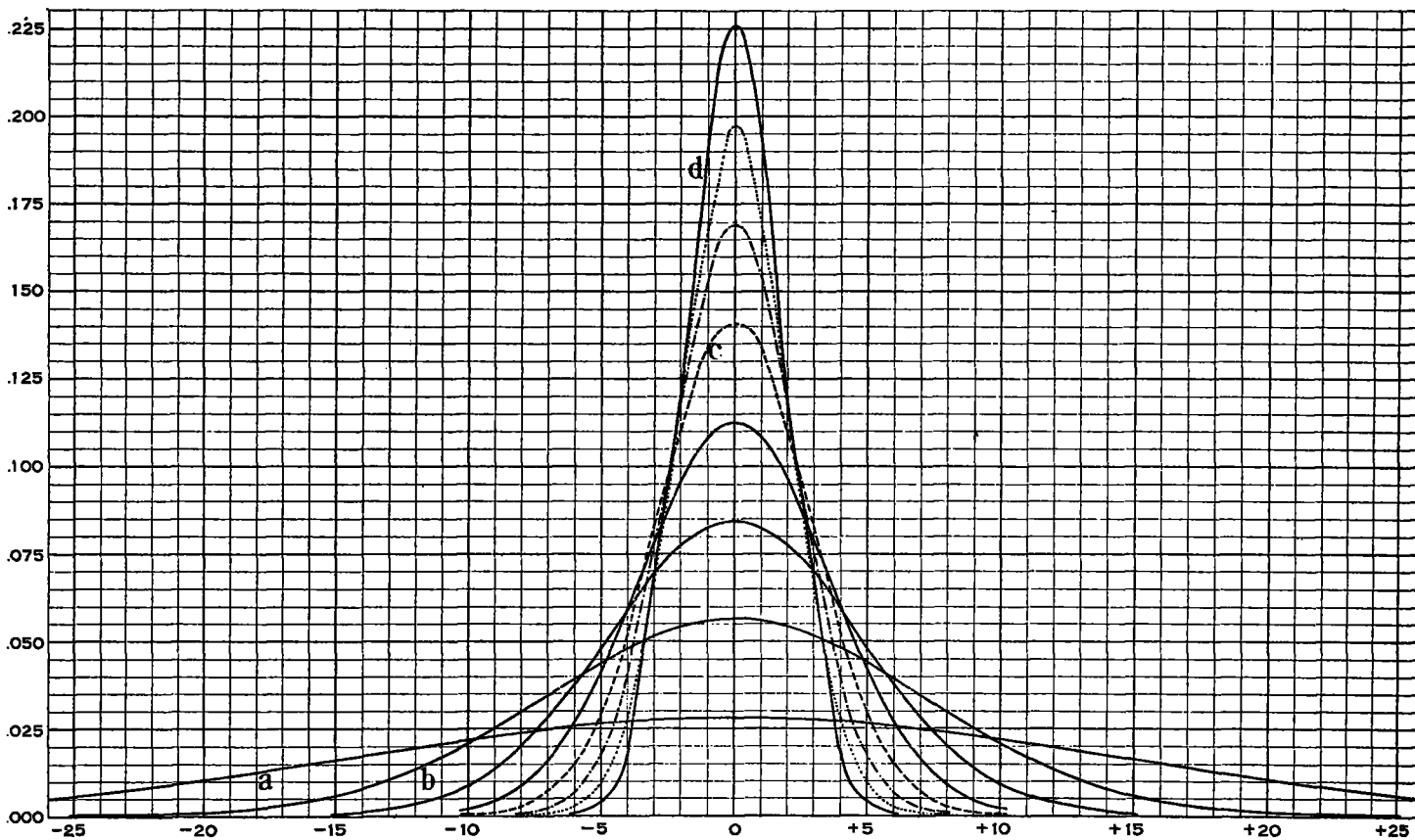
$0.005 \times 1 = 1/200$. Hence there will prove to be the equivalent of just 200 small rectangles under each curve.

Since the area of a curve represents the probability of all possible departures, so likewise the fractional area between ordinates represents the probability of departures lying between the limiting ordinates.

The probable error.

Mention has already been made that h in equation (1) is called the measure of precision. Its use, however, for this purpose seems to be confined almost exclusively to abstract mathematical discussions of equations of probability. Astronomers, physicists, and others engaged in precise measurements almost universally employ the *probable error* as a measure of precision or index of variability of their work.

The *probable error* or *probable variation* is that value of the departure that is just as likely to be exceeded

FIG. 7. Family of normal frequency curves on the type equation $y = (h/\sqrt{\pi}) e^{-h^2 x^2}$.

as not. It bears the same relations to the other errors that the *median* does to the original data; but the probable variation must not be confused with the median, as the two are wholly different. Half of the departures are equal to or greater than the probable departure and the remaining half have an equal or less value.

Every frequency distribution has two probable variations—one with a positive and the other with a negative sign. If the frequency distribution is normal then the numerical values of the two probable variations are identical. The probable variation ordinates $+E$, $-E$, in a frequency diagram, divide the area into three parts as in figure 8. The area between the ordinates is equal to half of the whole area and also equal to the sum of the areas of the remaining portions.

In skew or unsymmetrical distributions the probable variation as a value loses the significance assigned to it just above. In such a case the probable variation is simply the probable variation of a normal error curve of best fit to the particular group of skew data.

The probable variation is a measure of the amount of variation or disagreement in the values under discussion. When applied to measurements it is an index of accuracy. The smaller the probable error the greater the accuracy. When applied to data, the smaller the probable variation the smaller the dispersion or variation of the data from the mean.

Authorities give slightly different formulæ for computing the probable variation or probable error. *E. Hooker* (Quart. Jour., Roy. Met. Soc., 1908, 34: 281) gives the value of the probable variation as two-thirds the standard deviation. *Yule* (Theory of Statistics, pp. 306-307) places the value of the probable error at

$$E = 0.6745\sigma = 0.6745\sqrt{\frac{\sum x^2}{n}}$$

Davenport (Statistical Methods, 1915, p. 15) gives essentially the same value.

These formulæ may properly be applied when a large number of observations, n , is available; but the result is inaccurate for small values of n because the formula is deduced under mathematical assumptions that $\sum x^2$ represents the sum of squares for a large number of variations. When these requirements can not be regarded as satisfied, it is customary to use the following formula:

$$E = 0.6745\sqrt{\frac{\sum x^2}{n-1}}$$

(See Merriman: Method of least squares, 1915, p. 70.)

Astronomers, physicists, and others engaged in accurate measurements use the probable error frequently. Statisticians, however, and others engaged in like studies rarely use either the probable variation or the measure of precision as an index of variability, but employ a still different measure to indicate the characteristic deviations or dispersion of any given data. This measure is called the "standard deviation."

The standard deviation.

Those values of the departure which locate the points on a frequency curve where the curvature changes from convex to concave, that is, points of inflection, are values of the *standard deviation*. Like the probable variation, the standard deviation has positive and negative values which are numerically equal for *normal* distribution. The Greek letter small sigma (σ) is commonly employed to indicate standard deviation. When the

standard deviation is computed for an unsymmetrical distribution of data, its significance is only that of the standard deviation for the normal error curve of closest fit to the particular data. See $+\sigma$ and $-\sigma$ in figure 8.

The facility with which values of the standard deviation may be computed for a given set of data makes it a very convenient measure or index of variability and partly explains its adoption for this purpose.

Evaluation of h .—In equation (1) y is the probability (certainty being 1) that any given departure x will occur. In statistical work the *number* of departures of a given value, not its probability, are generally desired. The change in (1) necessary to obtain this result is easily effected. If y_n is the number of departures of a given value, x , and n is the total number of values or observations, then the probability of x is $y_n/n = y$, which substituted in (1) gives

$$y_n = \frac{nh}{\sqrt{\pi}} e^{-hx^2}. \quad (2)$$

In textbooks on the methods of least squares it is proved that for any given group of n variants conforming to the normal frequency, the value of h is given by the expression

$$h = \sqrt{\frac{n}{2\sum x^2}}, \quad (3)$$

in which $\sum x^2$ is the sum of the squares of all the departures from the mean.

It is also proven that the standard deviation, σ , is

$$\sigma = \sqrt{\frac{\sum x^2}{n}}. \quad (4)$$

The relations given in (3) and (4) introduced in (2) give:

$$y_n = \frac{n}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \quad (5)$$

This form of the equation in units of the standard deviation is the one commonly employed in statistical studies.

A still further simplification of the equation may be made. The expression $\frac{n}{\sigma\sqrt{2\pi}}$ is the value of the maximum ordinate, y_0 , that is, the value of y_n when $x=0$.

Let

$$y_0 = \frac{n}{\sigma\sqrt{2\pi}} \text{ and also let } X^2 = \frac{x^2}{\sigma^2},$$

then substituting these values in equation (5), the latter may be written

$$y_n/y_0 = Y = e^{-\frac{1}{2}X^2}.$$

This is the completely generalized equation of the normal frequency curve. The ordinates, Y , are measured in units of the maximum ordinate and the departures or abscissæ, X , are measured in units of the standard deviation. In this form the equation is applicable to every possible group of data conforming to the normal distribution. The logarithmic form of the equation is

$$\log Y = -\frac{\log e}{2} X^2 = -0.217145X^2.$$

Adding and subtracting 10 to the second member gives

$$\log Y = 10 - 0.217145X^2 - 10, \quad (6)$$

which puts the value of $\log Y$ derived from the equation

in the form in which logarithms are commonly tabulated. Figure 8 is the graphic representation of the equation (6).

Tables⁶ of values of Y obtained by this equation are to be found in many of the standard works on least squares, etc.

When $X=5$, Y is less than one part in 100,000. While the lateral branches of the curve extend to $\pm \infty$, the actual extension beyond $X = \pm 5$ is therefore almost inappreciable.

The following values serve to plot the curve shown in figure 8.

X	0	.2	.4	.5	1.0	1.5	2.0	3.0	4.0	5.0
Y	1.00000	0.98020	0.92312	0.88250	0.60653	0.32465	0.13534	0.01111	0.00031	0.00000

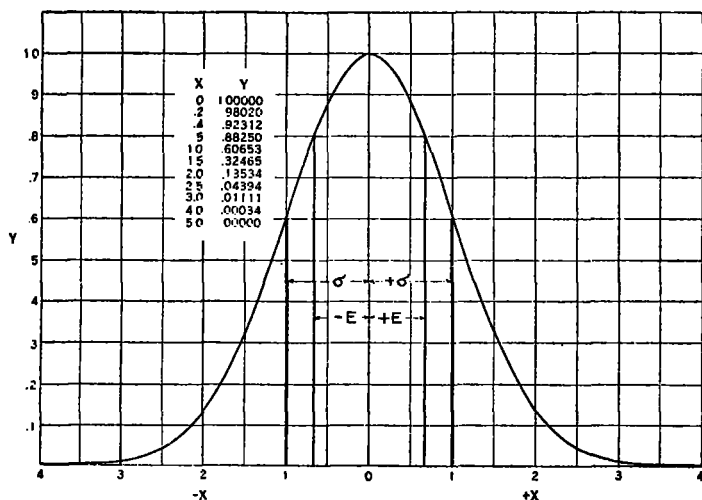


FIG. 8. Completely generalized normal frequency curve, $Y = e^{-\frac{1}{2}X^2}$.

The probability integral.—The methods of calculus provide means for evaluating the area of the generalized probability curve and tables of this area, otherwise known as the probability integral, have been prepared. Such tables⁶ in conjunction with tables of Y mentioned in the preceding paragraph are of great service in answering many important questions in probabilities.

Most probable departure; the mode.—In any frequency distribution that value of the variant which occurs most frequently is commonly called *the mode* (cf. the fashion). In the *normal* distribution, a zero departure is the most frequent, consequently the average value is the most frequent value. In the case of so-called skew variations or unsymmetrical frequency curves, however, the mode and the arithmetical mean differ more or less.

The median.—This name is applied to the middle value of a series of variants when arranged in order of their magnitude. Half of the values of the variant are equal to or greater and the remaining half are equal to or less than the median. In data conforming to the *normal* distribution the mean, the mode, and the median all have the same value. In the case of skew variations the three values differ more or less as indicated in figure 3 and the median in such cases always lies between the mean and the mode. According to Pearson⁷ the median is roughly one-third of the distance from the mean to the mode as measured on the axis of X .

PROCEDURE IN STATISTICAL STUDIES.

Data undergoing study and investigation should first be plotted, if possible, upon some appropriate system of coordinates. Plotting should be recognized as an indispensable step in the search for relations between observed phenomena supposed to be dependent on each other or similarly influenced by a third condition. The degree of interdependence, as well as the character of the relation (as mathematically defined), is indicated by the consistency with which the plotted data aligns itself to some recognized mathematical line or curve. Students are often content to stop when a smoothed, hand-drawn curve has been constructed to represent results, and often our meager knowledge of the problem affords no other course. In many cases, however, the mathematical law of relation is fully known, as well illustrated by the normal frequency curve, for example. In all such cases the work should not be regarded as finished until the equation of the best fitting line or curve has been evaluated. Such a mathematical curve is far superior to the smoothed, hand-drawn curve, for which the student has no better guide than the limited and probably imperfect data before him. The mathematical curve not only fits all the data the best possible, but defines the general law to which the observations are believed to conform more and more closely as their number and accuracy increases.

Moments of curves.—The mathematicians have borrowed from the physicists and have applied to the solution of statistical problems the idea of moments taken from mechanics in which the product of a given quantity (usually a force) by a distance from an axis or origin is called a moment. The *first moment*, commonly designated, v_1 , of a group of statistical data comprising n values is in reality simply the algebraic sum of all the departures from the mean or some convenient reference value, divided by n ; the *second moment*, v_2 , is the sum of the squares of the same departures divided by n ; the *third moment*, v_3 , is the sum of the cubes divided by n , and so on. Stated in this form the idea of moments as a product does not seem to enter. If, however, f , is the frequency, that is the number of departures, all of the same value x , then fx represents the sum of those departures and $\frac{\sum fx}{n} = v_1$ is the first moment of the data, similarly $\frac{\sum fx^2}{n} = v_2$ is the second moment, $\frac{\sum fx^3}{n} = v_3$ is the third moment, and so on. The analogy to moments is here clearly apparent.

Each mathematical curve also may be conceived to have moments similar to those above for the data. Such moments are commonly designated by the symbol μ and must be capable of calculation from the equation of the curve.

Mathematical curve fitting.—The completion of almost every study of statistical data, as well as of observations and measurements, leads finally to adjusting a particular mathematical curve represented by an equation to fit a given group of data. This requires the evaluation of the constants of the equation so as to secure the best possible fit. The methods of least squares provide certain definite processes for accomplishing this result by the formation and solution of so-called normal equations. These methods define the curve of *best possible fit* to be one for which the sum of the squares of the residuals (the observed values minus the computed values) is a minimum. The solution of the problem in some of the more complex and

⁶ Davenport, C. B. Statistical Methods, 1914. Tables III and IV.

Sheppard, W. F. New tables on the probability integral. *Biom.*, Feb. 1903, 2: 174.

⁷ Pearson, K. Variation of the egg of the sparrow. . . . *Biometrika*, 1, January, p. 261.

difficult cases by the methods of least squares is at times impracticable or even impossible, and Pearson⁸ has developed and applied the method of moments to curve fitting. By this method the area under the mathematical curve is equated to the area represented by some smooth curve passing through the observations. In addition, one or more successive moments of the curve (expressed in terms of the constants of the equation) are equated to the corresponding moments of the data. A sufficient number of simultaneous equations is thus obtained to evaluate the constants of the mathematical equation.

The methods of moments and of least squares are identical when employed in fitting parabolic curves of any order to observations, but the solution by moments in the case of exponential and other transcendental equations, a number of which are indispensable in statistical studies, offer important advantages. At the best the processes are complex and tedious, but must be mastered to a certain extent by those students in these fields of inquiry who would attain the highest results.

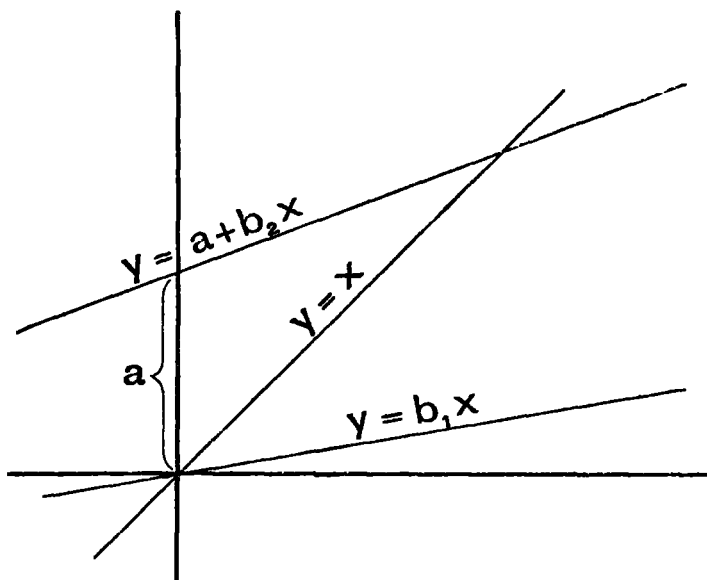


FIG. 9.—Illustrating straight lines and their equations.

A few of the multitude of mathematical equations that are likely to be employed to represent relations in statistical studies may be briefly noticed. Examples illustrating the least square methods of finding the most probable value of the constants in a few cases will be given later.

$$y_n = \frac{n}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \quad (7)$$

This apparently complex exponential equation for the normal frequency curve is very important and particularly interesting as it really contains but one undetermined constant. The most probable value of this may be very easily found for a given set of data.

$$\left. \begin{aligned} y &= x & (a) \\ y &= b_1x & (b) \\ y &= a + b_2x & (c) \end{aligned} \right\} \quad (8)$$

These are equations of straight lines—(a) through the origin inclined 45°, (b) through the origin inclined at

some other angle. Equations (a) and (b) are the kinds of equation obtained by calculations of the correlation coefficient, as will be shown later. Perfect correlation gives equation (a). Imperfect correlation leads to two equations of type (b), viz: $x = r \frac{\sigma_x}{\sigma_y} y$; $y = r \frac{\sigma_y}{\sigma_x} x$. Equation (c) is an inclined line cutting axis of y at distance a . (fig. 9.)

$$y = a + bx + cx^2 \quad (9)$$

The parabola has an equation representing a number of important physical laws and capable of being fitted to many groups of data, especially for limited ranges and where the law of relation does not involve points of inflection; that is, points at which the line changes its direction of curvature from concave to convex. Parabolic equations of the third and higher orders are frequently employed to represent more complex curves. In its general form the equation is

$$y = a + bx + cx^2 + dx^3 + \dots \quad (10)$$

This form of equation is better known as the Maclaurin expansion. The addition of higher powers no doubt increases the elasticity of the curve, so to speak, and enables it to adjust its curvature to fit the observations, but a careful inquiry to ascertain the real physical nature of the relations it is desired to represent may lead to a form of equation with fewer terms that gives a better fit. So much depends on the proper choice of equation that remarks by Pearson⁹ on this subject may be quoted here:

The hasty assumption of some physicists and many engineers that a parabola of the form

$$y = c_0 + c_1x + c_2x^2 + c_3x^3 + \dots$$

is always a good thing, is to be deprecated as may be seen at once by considering what a poor fit is obtained in this way to material really expressed by such curves as

$$y = y_0 e^{-cx^2}, \quad y = y_0 \sin nx, \quad y(x+c) = b^2, \text{ etc.}$$

To assume a curve of this form we must show the rapid convergency throughout the proposed range of the Maclaurin expansion, and this is not always feasible.

We find physicist and statistician remarking that "the increased accuracy of the result obtainable by least squares would not be an adequate return for the labor involved," and then falling back on some more or less questionable process of determining the constants. This process may be graphical or arithmetical, but it is usually unsystematic in character and elastic in result.

After fitting one of his skew frequency equations of the form $y = y_0 \left(1 + \frac{x}{a}\right)^p e^{-\frac{px}{a}}$, also a series of parabolas up to the 6th order, to a group of data he remarks:

The table compares these results with the successive parabolas up to the sixth and shows how a well selected curve with three constants can easily be superior to one with seven constants. This point is of special importance, for objections have been raised against the skew frequency curves just referred to on the ground that they give better fits than the normal curve because they have one or two more constants as the case may be. This is true; but they also give better fits than some other curves with double their number of constants!

$$\left. \begin{aligned} xy &= a^2 \\ xy &= a + bx + cy \end{aligned} \right\} \quad (11)$$

The curves represented by these equations are of hyperbolic type with infinite branches at right angles to each other. In the first, the branches are asymptotic to the axes (that is, approach and meet the axes at in-

⁸ Pearson, K. On the systematic fitting of curves to observations and measurements. *Biometrika*, April, 1902, 1:205.

⁹ *Biometrika* 1: 286, 267, 293.

finitely), in the second the asymptotes are at distances $\frac{a}{b}$ and $\frac{a}{c}$ from the axes.

$$y = ae^{bx} \quad (12)$$

This represents one of many forms of exponential equations. Its logarithmic form is,

$$\log y = a_1 + b_1 x \quad (13)$$

in which $a_1 = \log a$ and $b_1 = b \log e$.

Much careful attention must be given to finding the best values of the constants of such equations so as to properly fit the data. The textbooks should, but do not emphasize the fact that the least square methods ordinarily employed to determine the most probable values of a_1 and b_1 in equations like (13) may in the case of inexact data lead to an equation that fits the observations wretchedly.

The matter is too involved to discuss in the present paper.

$$y = a + b \sin nx + c \cos nx + \text{etc.} \quad (14)$$

Data which exhibits periodicity may be represented best by trigonometrical equations like (14). Additional terms with successive multiples of the angle may be added for more complex forms. While theoretically such an equation with sufficient terms can be made to represent any periodic curve whatsoever or fragment thereof, the work of harmonic analysis and curve fitting of complex curves like the tides, etc., can be successfully carried out only with elaborate instruments which have been invented for the purpose.

CORRELATION.

Having plotted his data the student may select from the several types of equations indicated in the foregoing the one he judges will best represent his problem and compute the constants thereof, thus securing the most definite and exact measure of relation possible between the quantities under consideration.

The procedure just described of plotting and curve fitting has to do with correlation in the highest sense of that word. Nevertheless, that definite procedure can not always be carried out satisfactorily, and the word correlation has come to be applied to a particular process of discovering interrelation of a limited character between statistical data in cases in which it may be uncertain even that any relation exists. Eagerness to utilize any mathematical process to aid in resolving the perplexities of the interrelations of groups of statistical data of all kinds may have led some, before examining the actual mathematical significance and basis of the coefficient of correlation, to form an exaggerated estimate of its value and possibilities. The following quotation from Yule¹⁰ indicates clearly the exact nature of the problem and the purpose served by the correlation coefficient.

The complete problem of the statistician like that of the physicist is to find formulae or equations which will suffice to describe approximately these curves.

In the general case this may be a difficult problem, but in the first place it often suffices, as already pointed out, to know merely whether on an average high values of the one variable show any tendency to be associated with high or with low values of the other, a purpose which will be served very fairly by fitting a straight line; and, further, in a large number of cases, it is found either (1) that the means of arrays lie very approximately round straight lines, or (2) that they lie so irregularly (possibly owing only to paucity of observations) that the

real nature of the curve is not clearly indicated, and a straight line will do almost as well as any more elaborate curve. In such cases—and they are relatively more frequent than might be supposed—the fitting of straight lines to the means of arrays determines all the most important characters of the distribution. We might fit such lines by a simple graphical method, plotting the points representing means of arrays on a diagram like those of figures 36–38, and “fitting” lines to them, say, by means of a stretched black thread shifted about till it appeared to run as near as might be to all the points. But such a method is hardly satisfactory, more especially if the points are somewhat scattered; it leaves too much room for guesswork and different observers obtain very different results. Some method is clearly required which will enable the observer to determine equations to the two lines for a given distribution, however irregularly the means may lie, as simply and definitely as he can calculate the means and standard deviations.

The *function of the correlation coefficient* is therefore clearly defined, namely, that of an index of the extent to which the relation between certain data may be represented by a straight line. A low correlation coefficient must not be interpreted to mean no relation necessarily, but that if a relation exists it is not well represented by a straight line.

In the search for a relation between two variables, three steps are possible: (1) To plot the data, selected and arranged in the manner to bring out best any relation that exists. This relation will be indicated by the alignment of the dots or points along some recognized line or curve. (2) We may proceed at once by least-square methods to find the equation of the plot, be it either a straight line or curve; or (3) concluding from the graph that a relation is apparent and that it can be represented by a straight line quite as well as by any curve, we may proceed at once to compute the correlation coefficient.

The first step is practically indispensable. When the second is performed the relation between the variables is defined and formulated in the most positive and complete manner. Nothing further is needed, for the reasons stated in the quotation from Yule. The correlation coefficient should, however, always be computed when the plot shows that a linear relation suffices and when for any reason the definite equation contemplated in the second step is not computed.

The correlation coefficient¹¹ is given by the equation

$$r = \frac{P}{\sigma_x \sigma_y} = \frac{\sum(xy)}{n \sigma_x \sigma_y}$$

$\frac{\sum(xy)}{n}$ is the sum of the products of the departures of x from its mean multiplied by the departure of the corresponding y from its mean—all divided by the number of pairs n . The expression σ_x is the standard deviation of the x departures; σ_y is the standard deviation of the y departures.

The following note by J. Warren Smith is of interest:

Value of the correlation coefficient.—In the past 10 years the writer has used the correlation coefficient as a practical method for showing the measure of the effect of the rainfall and temperature for definite periods and areas upon the yield of various crops.

The practice has been to first test the possible relation between the factors by means of the dot chart or curve chart or by some of the other recognized methods of sampling. During this time nearly 1,000 calculations have been made for the correlation coefficient, and no case is recalled when a low coefficient was found to be due to the law of relation being nonlinear. In every instance when using the dot chart, where there was a scattering of the dots, there was a low coefficient, and whenever there was a high value the dots were grouped around a fairly diagonal straight line. The dot chart may be made from the departure values or, what is still better, from the actual figures without the extra work of obtaining the departures from the normals.

¹⁰ Yule, G. U. Introduction to theory of statistics, p. 169.

¹¹ Yule, G. U. Introduction to the theory of statistics, 1910. p. 171, Equations (1) and (4).

Figure 10 is an example of the simplest method of determining by the dot chart whether the relation between two factors is linear. This chart shows the relation between the average rainfall for the States of Indiana, Illinois, Iowa, and Missouri during the month of July and the average yield of corn per acre. In this case it will be seen that there is a fairly regular increase in the corn yield with an increase in the rainfall, and while there may be a slight tendency to a decrease in yield, with the greatest rainfalls, nevertheless this is not significant. My value for the correlation coefficient was $+0.61 \pm 0.08$.

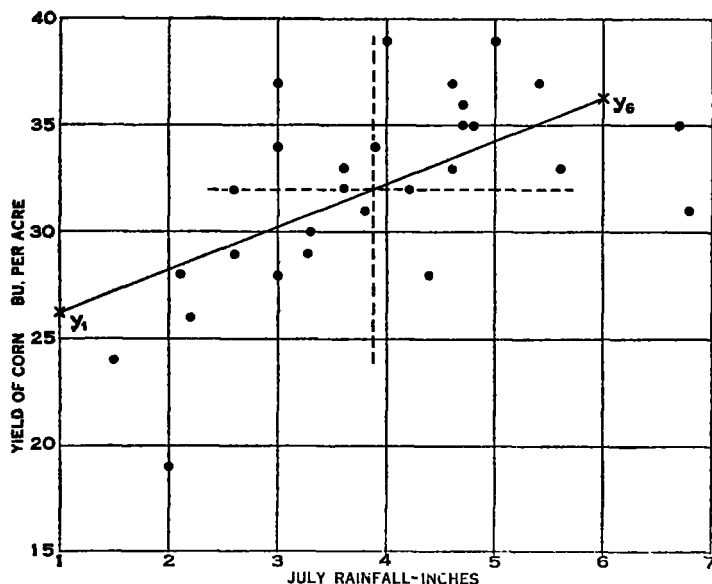


FIG. 10. Dot chart of data on the rainfall and the yield of corn in the States of Iowa, Missouri, Illinois, and Indiana. Also straight line of closest fit.

Figure 11 indicates the position of the dots when the line of closest fit is nearly parallel with one of the axes—in this case, with the axis of abscissæ. The correlation coefficient from this calculation is only -0.14 , thus agreeing with the promiscuous scattering of the dots.

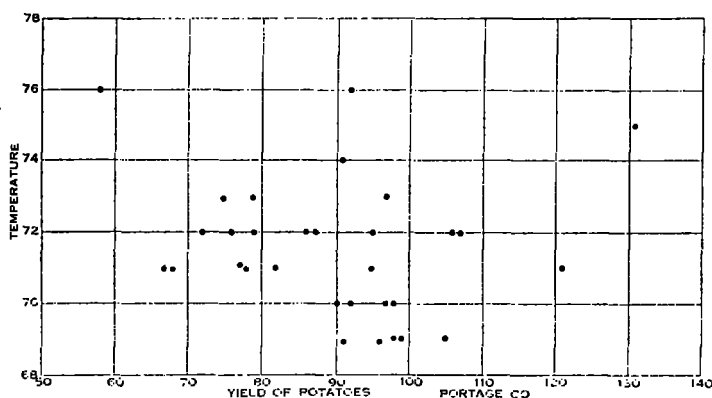


FIG. 11. Dot chart of yield of potatoes and temperature indicating little or no relation because straight line of approximate fit is nearly or quite parallel to one axis.

Figure 12, on the other hand, shows at once that one may expect a high correlation coefficient value, but instead of a straight diagonal line from the intersection of the axes being the closest fit, the line runs from the extremes of the axes. This shows that the correlation coefficient will have a minus sign, because as one factor increases the other decreases, instead of both increasing together as in figure 10. This figure shows the relation between the mean temperature for the month of July and the average yield of potatoes for the State of Ohio. * * *

The writer began the use of this method of studying the relation between weather and crop yields as an experiment, not knowing that it had ever before been used for this purpose. We have since then, however, learned that Mr. G. Udny Yule and Mr. R. Hooker had already recognized its value in this connection and had used it extensively. We wish to strongly urge the continuation of the use of the correlation table in studying weather and crops, with proper preliminary examination, until some better method has been evolved.—J. Warren Smith.

Emphasis has been laid on the necessity of limiting the application of the methods of correlation only to data in

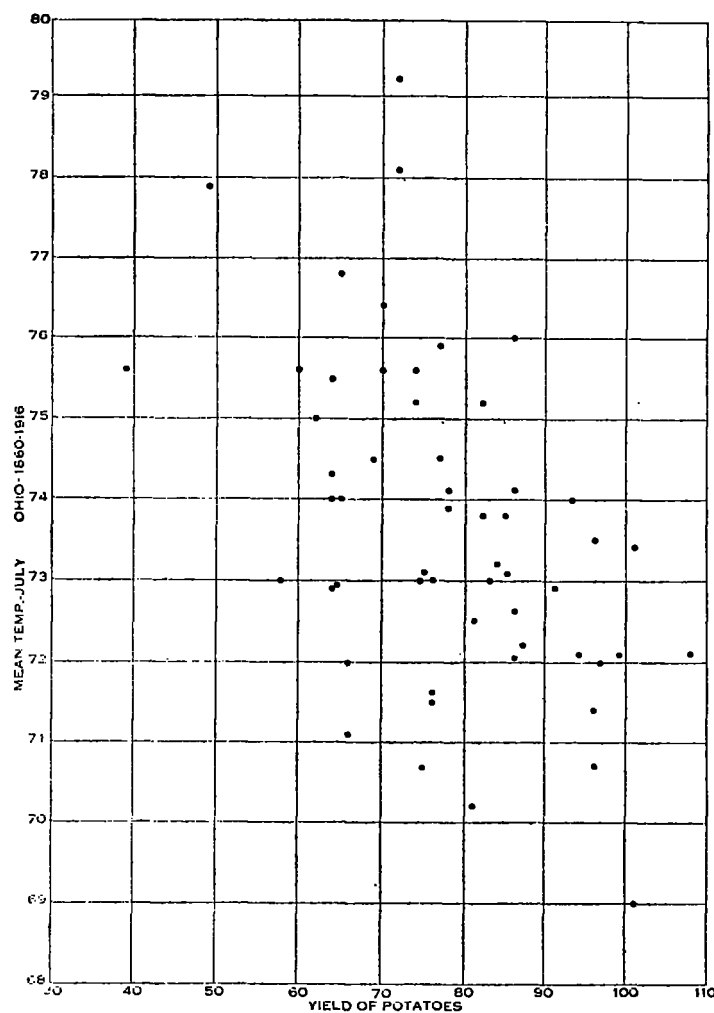


FIG. 12. Dot chart of yield of potatoes and temperature in which line of approximate relation inclines to the left.

linear relation. The following values of x and y representing points upon a parabola were submitted without explanation, to have the coefficient of correlation computed.

x	19.6	15.9	19.2	15.0	13.9	18.8	8.8	17.7	10.2	20.0	12.8
y	33	39	26	20	41	35	45	24	16	30	18

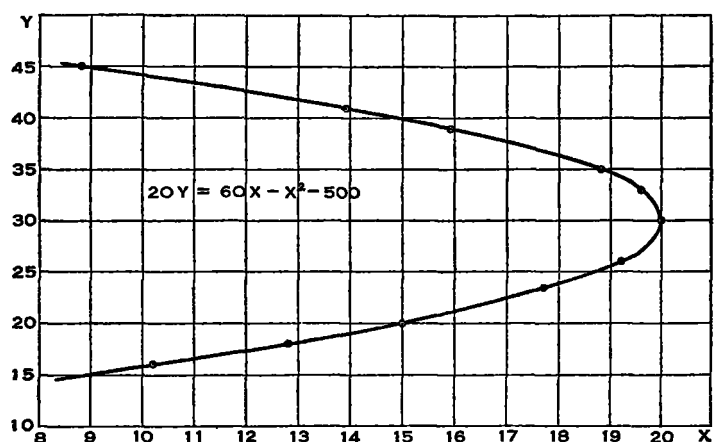


FIG. 13. Dots on parabolic curve. Coefficient of correlation is very nearly zero, because data can not be represented by a straight line.

The correlation coefficient was found to be -0.0048 showing no relation. The points are shown in figure 13.

A still more interesting case has been worked out by Mr. W. G. Reed, of the United States Office of Farm Management, who has analyzed data on the tides and phases of the moon.

A coefficient of correlation may be near zero when there is very close relationship, as is shown in such a condition as the relationship between the height of high water and the phase of the moon which is shown for Old Point Comfort, Va., figure 14. The figure indicates that the relation is harmonic. Although there is a close and very definite relation between the phenomena, the coefficient of correlation is near zero (0.106 ± 0.088) because the different portions of the curve of relation are such that a straight line along an axis will most nearly satisfy all the points. Of course, the angle is then zero and its tangent is zero.

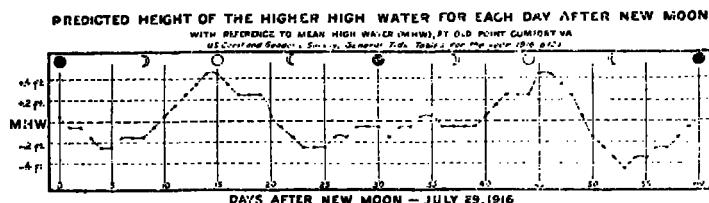


FIG. 14. Tidal data and phases of the moon with very definite relation but low value of coefficient of correlation.

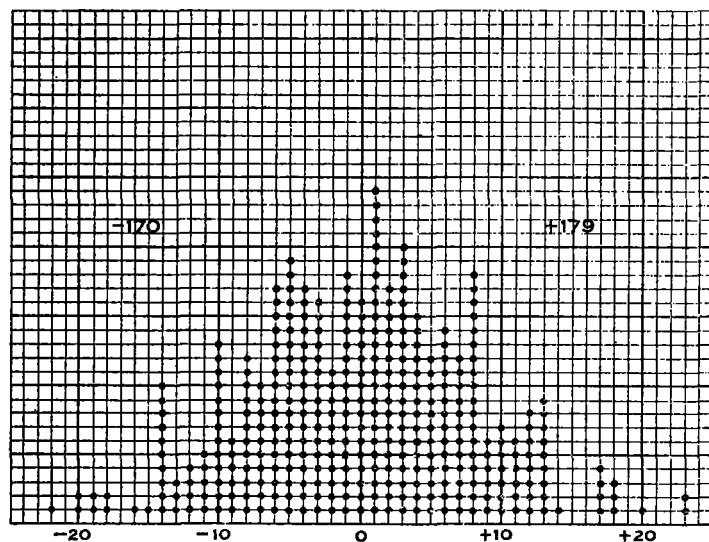


FIG. 15. Dot chart of frequency of departure of July daily mean temperature from normal for many years for Springfield, Ill.

These illustrations are given to show the importance of the proper use of the correlation method of analysis and should not be construed by the reader as discrediting the value of the principle when employed in a legitimate and intelligent manner.

EXAMPLES.

The examples given below are selected to illustrate the practical application of the principles presented in the foregoing and give the student a fuller idea of the details of solving actual problems.

Example 1.

Normal frequency curve representing departure of daily mean temperatures from the normal or average for the day. Data selected from observations at Springfield, Ill., and Fresno, Cal., for months where the average for the month was very nearly normal.

Figures 15 and 16 illustrate by dots the departures from the normal for each day for a year's observations at

each station. The accuracy of the data is not now a question, as its present purpose is to serve simply as an example.

The relatively small amount of data as well as the inexact manner of its compilation and the tendency of climatic data to exhibit wide fluctuations cause marked irregularities and abnormalities to appear in these frequency polygons, even when several hundred observations are available. The data for Springfield show more dispersion than those for Fresno, with a slight excess of positive departures. The negative departures are in excess for Fresno and there the departures show noticeably less dispersion. More observations are required to determine whether or not these features are real characteristics of the climate of the two stations or are due to errors from dropping decimals and other accidental causes incident to the method employed in forming these particular departures. These matters, however, need not concern us now, as our present object is to illustrate methods rather than establish definite facts.

We, therefore, simply combine the two sets of data into one set by addition, obtaining the results shown in Table 2 and figure 17. In spite of some lack of symmetry and a few marked abnormalities we regard the frequency polygon as quite satisfactory and characteristic. Assuming that a mathematical equation like (1) best represents the law of frequency of these data, it remains only to find the equation of the curve of best fit; that is, to find the value of h in (2) or σ in (5) corresponding to the 731

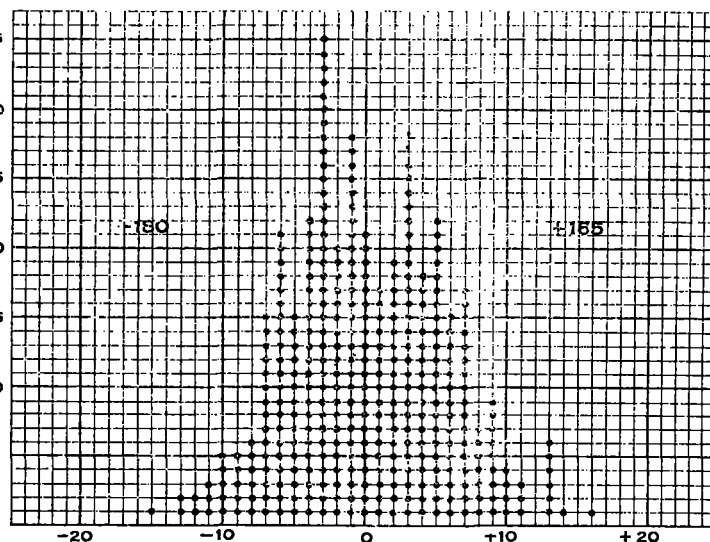


FIG. 16. Dot chart of frequency of departure of July daily mean temperature from normal for many years for Fresno, Cal.

values of the temperature departure data. Fortunately this is done very easily, by simply calculating σ from the standard deviation equation (4).

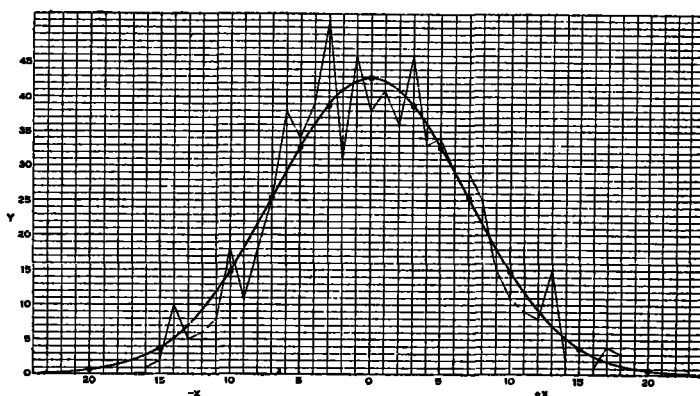


FIG. 17. Combined frequencies of Springfield-Fresno temperature departures and the normal frequency curve of best fit. (See equation 16.)

Practical calculation of the standard deviation.—Table 2 gives the data and steps in the computation in full. The calculation is such an important one on account of its frequent occurrence in physical and statistical work that it justifies special notice.

In equation (4) on page 557, Sigma x^2 (Σx^2) must represent the *least sum of squares*. As already explained this sum will be the least only when the departures are taken with respect to the true mean.

Now, ordinarily, this kind of a departure can not be tabulated, *accurately*, because the mean value generally ends in a long decimal which it is impracticable to retain in the computations. Nevertheless whatever part of the mean is rejected causes the sum of the squares to be too large. Fortunately the exact value of the excess can be easily found so that in practice it is more convenient to select some arbitrary number, generally near the mean, and form the departures with respect to this value. The minimum sum of squares can then be easily found, as also the exact value of the mean.

Obviously, we need some convenient nomenclature to indicate to the eye whether, for example, a given sum of squares, as also other related quantities, is based on departures from the true mean or from some arbitrary reference number. We propose to adopt the following:

Σx , Σx^2 , Σxy , etc., designate the sums of the quantities represented by x , x^2 , xy , etc., when departures are calculated from the true mean.

$[x]$, $[x^2]$, $[xy]$, will indicate the sums of departures taken from some arbitrary number.

We now need to know the relation between Σx^2 and $[x^2]$. According to Yule¹² and other authorities this relation is indicated in the customary equation for the standard deviation, which may be written thus in our notation:

$$\sigma = \sqrt{\frac{[x^2] - nd^2}{n}}$$

in which d is the difference between the true mean and the base or reference number used in forming the departures. While mathematically exact this form of the equation for σ is faulty and troublesome from the computer's point of view, because the term d^2 can not be accurately computed by squaring d unless the latter is computed with more significant figures than is otherwise necessary. The remedy for this is found by stating the equation in the following identical form:

$$\sigma = \sqrt{\frac{[x^2] - \frac{[x]^2}{n}}{n}}$$

where $[x]^2$ is the algebraic sum of the departures *squared*. Ordinarily this sum will be computed any way as a check on the work. It will generally be a small number whose square can be exactly computed, as also the quotient $[x]^2/n$, thus giving with the least amount of work the minimum sum of squares $\Sigma x^2 = [x^2] - [x]^2/n$. All the work is shown in Table 2. If M is the approximate mean or base number, the true mean A is found from the following equation:

$$A = M - [x]/n$$

In the present example the quantity A has no special significance in connection with the temperature data because of the manner in which the departures were taken. It does, however, indicate that the normal curve of best

fit is shifted by 0.0506° toward the positive side of the axis Y , and that the departures were taken from very nearly the true mean.

TABLE 2.—Fresno-Springfield temperature departure data.

[Degrees Fahrenheit.]

Departures, x .	Number, y .	x^2 .	xy .	x^2y .
-22.....	1	484	-22	484
-20.....	2	400	-40	800
-19.....	2	361	-38	721
-18.....	2	324	-36	648
-16.....	1	256	-16	256
-15.....	2	225	-30	450
-14.....	10	196	-140	1960
-13.....	5	169	-65	845
-12.....	6	144	-72	864
-11.....	8	121	-88	968
-10.....	18	100	-180	1800
-9.....	11	81	-99	891
-8.....	18	64	-144	1152
-7.....	25	49	-175	1225
-6.....	38	36	-228	1368
-5.....	34	25	-170	850
-4.....	39	16	-156	624
-3.....	51	9	-153	459
-2.....	31	4	-62	124
-1.....	46	1	-46	46
± 0.....	37			
+ 1.....	41	1	+41	41
2.....	36	4	72	144
3.....	46	9	138	414
4.....	33	16	132	528
5.....	34	25	170	850
6.....	29	36	174	1044
7.....	29	49	203	1421
8.....	25	64	200	1600
9.....	15	81	135	1215
10.....	11	100	110	1100
11.....	9	121	99	1081
12.....	8	144	91	1152
13.....	15	169	195	2535
14.....	2	196	28	392
16.....	1	256	16	256
17.....	4	289	68	1156
18.....	3	324	54	972
20.....	1	400	20	400
23.....	2	529	46	1058
Positive departures.....	344		1997	33894
Zero departures.....	37			
Negative departures.....	350		-1960	
Total.....	$n=[y]=731$		$[xy]=+37$	$[x^2y]=33894$

$$\frac{[xy]}{n} = +0.0506.$$

$$\text{Minimum sum of squares} = 33894 - \frac{(37)^2}{n} = 33892.13.$$

$$\text{Standard deviation} = \sigma = \sqrt{\frac{33892.13}{731}} = 6.809.$$

The final result of the analysis of the Springfield-Fresno temperature data, assuming that the same law applies to both stations, is the frequency law given by the following

$$\sigma \text{ (standard deviation)} = 6.809^\circ \text{ F.}$$

and by substituting σ in (5) we have

$$y \text{ (frequency of departures)} = 42.83e^{-0.010783x^2}. \quad (15)$$

For easy calculation of values of y the equation (15) may be put in the form

$$\log y = 1.6374 - 0.004683x^2 \quad (16)$$

whence the following values of x and y are derived:

x	0	3	5	7	10	15	20	25
y	42.8	38.9	32.7	25.2	14.7	3.8	0.6	0.05

We wish to know how much confidence may be placed in these final results. This is ascertained by computing

¹² Yule, G. U. Introduction to the theory of statistics. p. 135, equation (4) transposed.

the probable error¹³ of the standard deviation, which turns out $\sigma = 6.809^\circ \pm 0.12^\circ$. This indicates that when the departures do not differ greatly from the standard deviation the results obtained from (15) or (16) are not likely to be more than $\pm 2\%$ in error.

The student must catch the significance of the values of y computed by means of (16). According to these data the temperatures will be the same as the normal for the day on 42.8 days out of 731, i. e., a period of two years (one a leap year). It will go 3 degrees above or below the normal 39 times. The departure will be ± 15 degrees on less than 4 days in these two years. Half of all the departures will be above and half will be below the value of the probable departure¹⁴ which will be $0.6745\sigma = 4.59^\circ$. The table shows 360 out of 731 values between ± 4 and -4 degrees, a slight excess over the number called for by the curve, but still a close agreement. Suppose we wish to know the percentage of departures that will exceed, say 10, 15, etc. degrees? Questions of this kind are readily answered by reference to a table of the probability integral. Davenport's Table IV gives, for the entry $10/\sigma = 1.47^\circ$, the value 0.42922, which represents the half area of the probability curve out to $x = +10$ degrees. The area of half the curve is 0.50000, therefore the portion beyond $x = +10$ will be 0.071. An equal extension will lie beyond $x = -10$, hence the percentage of departures beyond ± 10 degrees will be $2 \times 0.071 = 14\%$. The percentage beyond 15 degrees will be 2.8%. These results indicate that about 11% of the departures will have values between 10 and 15 degrees.

The foregoing can not be stated as positive facts concerning Fresno and Springfield climates, because the data were not sufficiently representative to justify positive statements, but the examples serve to illustrate how such results can be obtained by accurate mathematical methods. Moreover, we have the satisfaction of knowing that frequency curves, when properly deduced in the manner indicated even from seemingly irregular and insufficient data, nevertheless express a definite law of occurrence of deviations that not only fits the data employed the best possible, but is the general law to which the data will conform more and more closely as the length of the records is prolonged and the number of observations multiplied.

Problem II.

To find the constants of the equation of the straight line of best fit to the data in Prof. Smith's example, figure 10, giving the relation between July rainfall for the States of Indiana, Illinois, Iowa, and Missouri. Let y = yield of corn per acre for this territory and r the July rainfall. These data give a series of observation equations for a straight line of the type

$$y = a + br,$$

and our problem is to determine the best or most probable values of a and b . The data and computations are given in Table 3, and figure 10 shows the data in graphic form. A few of the observation equations may be written thus:

$$34 = a + 3.9b$$

$$33 = a + 4.6b$$

$$26 = a + 2.2b$$

$$* * * * *$$

$$31 = a + 6.8b$$

There will be in all 28 equations and, according to the methods of least squares, the most probable values of a and b are given by the solution of two normal equations.¹⁵

Rule.—The normal equation for a is formed by multiplying each observation equation by the coefficient of a (with its proper sign) in that equation. The sum of the resulting equations is the normal equation for a .

Likewise the normal equation for b is formed by multiplying each observation equation by the coefficient of b in that equation. The sum of the resulting equations is the normal equation for b .

The two normal equations thus obtained are simultaneous equations and their solution gives the most probable values of a and b .

Carrying out the operations indicated the normal equations may be written in this form—

$$\begin{aligned}\Sigma y &= na + b\Sigma r \\ \Sigma ry &= a\Sigma r + b\Sigma r^2\end{aligned}$$

These equations give

$$b = \frac{n(\Sigma ry) - (\Sigma r)(\Sigma y)}{n(\Sigma r^2) - (\Sigma r)^2} = 2.027$$

When b is found,

$$a = \frac{\Sigma y - b(\Sigma r)}{n} = 24.07$$

In these equations Σr is the sum of the rainfall and Σy the sum of the yields. Σry is the sum of the products, rainfall by yield, and Σr^2 is the sum of the squares of the rainfall. These values are similar to quantities that require to be computed in forming the correlation coefficient and the non-mathematical student does not need to know how to deduce all the equations, but simply to be able to perform the computations shown in Table 3.

TABLE 3.—July rainfall and yield of corn for the States of Iowa, Missouri, Illinois, and Indiana.

Year.	Rain, r .	Yield, y .	r^2	ry	Computation.
	<i>Inches.</i>	<i>Bu. p. ac.</i>			
1888.....	3.9	34	15.21	132.6	
1889.....	4.6	33	21.16	151.8	
1890.....	2.2	26	4.84	57.2	
1891.....	3.0	34	9.00	102.0	
1892.....	4.4	28	19.36	123.2	
1893.....	3.0	28	9.00	84.0	
1894.....	1.5	24	2.25	36.0	
1895.....	4.7	35	22.09	164.5	$\Sigma r = 109.0$ $(\Sigma r)^2 = 11,881$ $\Sigma y = 895$
1896.....	6.7	35	44.89	234.5	$\Sigma r^2 = 472.56$
1897.....	3.3	29	10.89	95.7	$\Sigma ry = 3,581.9$
1898.....	3.6	32	12.96	115.2	
1899.....	3.6	33	12.96	118.8	
1900.....	4.8	35	23.04	168.0	$(\Sigma r)(\Sigma y) = 97,555$ $n(\Sigma r^2) = 13,231.68$
1901.....	2.0	19	4.00	38.0	$n(\Sigma ry) = 100,293.2$
1902.....	5.4	37	29.16	199.8	
1903.....	3.8	31	14.44	117.8	
1904.....	4.2	32	17.64	134.4	$b = \frac{100,293.2 - 97,555}{13,231.68 - 11,881}$ $= \frac{2,738.2}{1,350.68} = 2.027$
1905.....	5.0	38	25.00	190.0	$a = \frac{895 - 231.943}{28} = \frac{674.057}{28}$ $= 24.07$
1906.....	3.0	37	9.00	111.0	
1907.....	5.6	33	31.36	184.8	
1908.....	3.3	30	10.89	99.0	
1909.....	4.7	36	22.09	169.2	
1910.....	4.6	37	21.16	170.2	
1911.....	2.6	32	6.76	83.2	
1912.....	4.0	39	16.00	156.0	
1913.....	2.6	29	6.76	75.4	
1914.....	2.1	28	4.41	58.8	
1915.....	6.8	31	46.24	210.8	
Sums.....	109.0	895	472.56	3,581.9	

¹³ Davenport, C. B. Statistical methods, 1914, p. 16. $E_\sigma = \pm 0.6745\sigma/\sqrt{n}$.

¹⁴ See above under "Probable error," p. 557.

¹⁵ Merriman. Methods of least squares. New York, 1915. par. 48.
Comstock. Method of least squares. Boston, 1899. pp. 19 and 23.

The equation of the straight line that best fits the data of Table 3 and represents the relation between the yield of corn and the July rainfall is

$$y = 24.07 + 2.03r.$$

This equation indicates that if r is 1, $y = 24.07 + 2.03 = 26$. If $r = 6$, $y = 24.07 + 12.16 = 36$. Two points like these suffice to locate the line of best fit on the diagram, figure 10, as at y_1 and y_6 .

The coefficient $+2.03$ means that each inch of increase in July rainfall will add 2.03 bushels per acre to the yield of corn in the States considered. It must be understood, of course, that the law of relation represented by the equation is purely an arbitrary one and applies only to conditions *within the range of the records discussed*. Moreover the relation is only approximate, as is shown by the comparatively widely scattered distribution of the dots in figure 10. Obviously other factors than rainfall influence the yield and should be considered, but the line found by the foregoing method is, in a sense, a first approximation in the discovery of the relations between the statistical data considered and is identical, although expressed in different units of measurements, with the line defined by the correlation coefficient, $r = 0.61$, previously found.

Problem III.

Variation in December minimum temperatures at Washington, D. C., 1872 to 1915, inclusive. These data afford an instructive illustration of the application of the methods of statistics to climatology. From 1872 to 1888

the observations were made at the old office near 17th and G streets NW., and after 1889 at the present office at 24th and M streets NW. The data were first analyzed in two groups corresponding to the locations of the office; and finally treated as a whole independent of location, since there is no evidence of any systematic difference accompanying the removal of the office.

Table 4 gives the number of times (y_n) that the various minimum temperatures were observed in the whole period of 44 years, as also for the separate intervals of 17 years in the old location and 27 years in the present position.

In figure 18 the three groups of data are shown reduced to a 10-year basis of record, including the normal frequency curve of best fit for the whole series. The separate curves for the two groups of data, old and new office, differ too little from the one for the whole period to justify drawing them in, but the corresponding ordinates for all the curves are given in Table 4.

TABLE 4.—Computed values of y_n for temperature data, 10-year basis.

	Old office.	New office.	Combined.
x	y_n	y_n	y_n
± 0	13.31	14.52	13.93
± 2	13.00	14.12	13.53
± 4	12.13	13.00	12.58
± 6	11.52	12.22	11.88
± 7	10.02	10.36	10.21
± 10	7.46	7.29	7.39
± 15	3.62	3.08	3.34
± 20	1.31	0.92	1.10
± 25	0.36	0.20	0.28
± 30	0.07	0.03	0.05
± 35	0.1	0.003	0.01

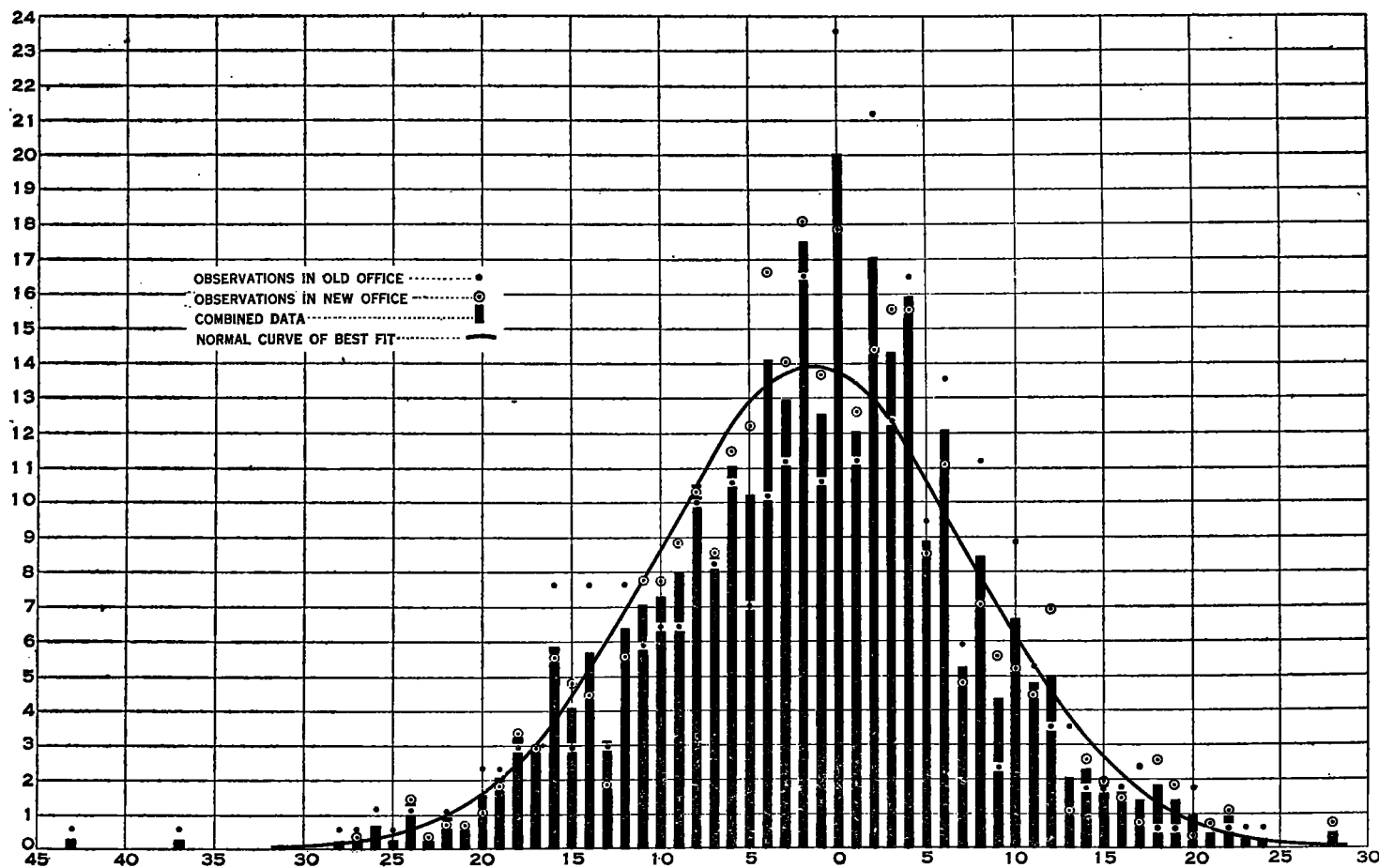


FIG. 18. The departures, above and below 30° F., of minimum temperatures of Washington, D. C., for December, 1872 to 1915. Also normal frequency curve of best fit.

The results of the calculation of the frequency curves, presented in Table 5, show the close agreement in the data.

TABLE 5.

Series.	Average minimum temperature, December.	Standard deviation, σ .	Probable variation, $\frac{1}{2}\sigma$.	Preponderance of even temperatures.		
				Even temperature.	Odd temperature.	Per cent.
Old office, 17 years.....	28.6	9.29	6.2	325	201	1.62
New office, 27 years.....	28.6	8.52	5.7	466	371	1.26
Whole series, 44 years.....	28.6	8.81	5.9	792	571	1.39

Preponderance of even temperatures.—A striking abnormality consisting of the preponderance of *even* temperature values is obvious from an inspection of figure 18. This is likely to be characteristic of a large proportion of meteorological data from which decimal fractions or unnecessary digits have been disposed of according to the customary rule.¹⁶ The minimum temperatures in the present case were read and recorded to the nearest tenth of a degree, and the fractions were subsequently disposed of according to the well-known rule which changes all readings ending in 5 to a number ending in an *even* digit. The effect of this rule upon the frequency distribution is to cause all observed temperatures between, say 9.5° and 10.5°, to be classified under 10°, and similarly, for any other *even* value, whereas any odd temperature like 7° comprises only the values between 6.6° and 7.4°. The latter, that is the odd class, comprises nine values, whereas the even class comprises 11 possible values. On the strict theoretical basis there should be about 22 per cent more even readings than odd in a large group of data subject to this cause of abnormality.

The last three columns in Table 5 show that the preponderance of *even* values considerably exceeds the theoretical expectation. The probable explanation of this excess is to be found in the fact that when the observer should read the fraction 0.4 or 0.6 he nevertheless is prone to take off and record a fraction of 0.5. This proneness of the observer to give preference to readings ending in 0.5 also extends to readings ending in zero tenths. No serious error of any kind arises from this more or less inherent and systematic abnormality except to give a distinct preponderance of *even* values. The cooperative observers of the Weather Bureau read and record minimum temperatures only to the nearest whole degree, so that excess of even values should not be expected in such records. The December observations for 20 years at a cooperative station near Washington, have been tabulated and it is found there are 307 even values and 303 odd ones. Ten readings were missing. The approach to equality is here quite as close as can be expected.

A method of equalizing the distribution and eliminating the abnormality due to the preponderance of *even* values is indicated in Table 6. The equations of adjustment are:

$$\begin{aligned} B &= a + 2b + c, \\ C &= b + 2c + d, \\ D &= c + 2d + e, \\ &\text{etc.} \end{aligned}$$

¹⁶ The Weather Bureau rule for disposing of decimals is given in its "Instructions for preparing meteorological forms," § 123, and reads as follows:

"123. In dropping decimals, if the figure to be dropped is greater than 5 (or 5 with a remainder) the preceding figure will be increased by 1. If the decimal figure to be dropped is 5 exactly, the preceding figure when odd will be increased by 1, and when even it will remain unchanged. If the figure to be dropped is less than 5, retain the preceding figure unchanged."

TABLE 6.—Illustrating elimination of abnormality due to preponderance of even values of data.

Class.	Range in class.	Frequency.	
		Observed.	Equalized.
4	11	a = 0	A = 0
5	9	b = 0	B = 8
6	11	c = 8	C = 18
7	9	d = 2	D = 27
8	11	e = 15	E = 37
9	9	f = 5	F = 32
10	11	g = 7	G = 19
11	9	h = 0	H = 7
12	11	i = 0	I = 0

The equalized frequencies each comprise 40 elements, or we may regard them each as four times the true adjusted values. Fractions will be introduced, however, if we divide by 4, and for easy and accurate computation it will generally be best to use the adjusted values directly and divide by 4 at the end of the computation, if desired.

The whole series of December minimum temperatures for Washington were equalized for odd and even values by the rule mentioned above with the result:

	Even.	Odd.	Total.
Data as tabulated.....	792	571	1363
Data as adjusted.....	2726	2726	5452

Probability of a given departure.—Among the useful deductions to be drawn from the mathematical analysis of the minimum temperature data of Problem III are the probabilities of occurrence of given departures, or departures between certain limits. Some such results are given in Table 7, which is constructed from tables of the probability integral.

TABLE 7.—Number of days in 100 years on which the December minimum temperature at Washington, D. C., will differ from 28.6° by certain amounts.

[Based on 44 years' record. Standard deviation $\sigma = 8.88^\circ$.]

Departure.	$\leq \pm 5^\circ$ or less.	$\leq \pm 5^\circ$.	$\leq \pm 10^\circ$.	$\leq \pm 15^\circ$.	$\leq \pm 20^\circ$.	$\leq \pm 25^\circ$.	$\leq \pm 30^\circ$.	$\leq \pm 35^\circ$.	$\leq \pm 40^\circ$.
σ/σ	0.503	0.503	1.126	1.689	2.252	2.815	3.378	3.941	4.505
Number of days.....	1,322	1,778	807	283	75.4	15.1	2.26	0.25	0.02

An examination of the original data shows that no plus departure in excess of 30° has occurred, but, on the other hand, two negative departures exceeding 30° have occurred in the 44 years comprised by the record. Furthermore, a critical examination of figure 18 clearly indicates that the positive and negative departures are not equal in number or strictly symmetrical in arrangement. For the best results, therefore, we require an unsymmetrical type of curve, such as some of the types proposed by Pearson. Mr. Howard R. Tolley, of the Office of Farm Management, has recently developed a relatively simple modification of the normal frequency equation not yet described and published,¹⁷ that is appropriate for representing the moderately unsymmetrical or skew frequency distributions likely to be found in the analysis of nearly all climatic data.

The discussion of unsymmetrical distributions, however, can not be included in the present paper and must be reserved for later presentation. It may be appropriate, however, to point out here that systematic and consistent lack of symmetry of a frequency distribution indi-

¹⁷ To appear in this REVIEW for November, 1916.

cates that some one or more influences are acting which tend to make deviations of a particular kind or class to preponderate. The thing to do, therefore, is to discover and eliminate the preponderating influence. This procedure should ultimately suffice to reduce all distributions simply to the normal error curve of best fit.

Example IV.

Practical calculation of coefficient of correlation, using data in Example II on the relation between July rainfall for the States of Indiana, Illinois, Iowa, and Missouri, and the yield of corn per acre.

Table 8 gives the data and the computations in full. In the case of the rainfall the departures are computed from the arbitrary number 4.0, and 32 is an arbitrary base number used in tabulating the variations in yield of corn.

As already explained on page 563 in the practical calculation of the standard deviation, the use of the arbitrary base numbers causes the sums of the squares to be too large. The sum of the products represented by xy is also too large for the same reason. It was shown that the minimum sum of squares is given by the equation

$$\Sigma x^2 = [x^2] - \frac{[x]^2}{n}.$$

TABLE 8.—Coefficient of correlation between July rainfall in Indiana, Illinois, Iowa, and Missouri and the yield of corn per acre.

[By J. Warren Smith.]

Rainfall.				Yield of corn.			
Year.	r	Departure, x	x^2	Y	Departure, y	y^2	xy
	Inches.			Bushels.			
1888.....	3.9	- 0.1	0.01	34	+ 2	4	- 0.2
1889.....	4.6	+ 0.6	0.36	33	+ 1	1	+ 0.6
1890.....	2.2	- 1.8	3.24	26	- 6	36	+ 10.8
1891.....	3.0	- 1.0	1.00	34	+ 2	4	- 2.0
1892.....	4.4	+ 0.4	0.16	28	- 4	16	- 1.6
1893.....	3.0	- 1.0	1.00	28	- 4	16	+ 4.0
1894.....	1.5	- 2.5	6.25	24	- 8	64	+ 20.0
1895.....	4.7	+ 0.7	0.49	35	+ 3	9	+ 2.1
1896.....	6.7	+ 2.7	7.29	35	+ 3	9	+ 8.1
1897.....	3.3	- 0.7	0.49	29	- 3	9	+ 2.1
1898.....	3.6	- 0.4	0.16	32	+ 0	0
1899.....	3.6	- 0.4	0.16	33	+ 1	1	- 0.4
1900.....	4.8	+ 0.8	0.64	35	+ 3	9	+ 2.4
1901.....	2.0	- 2.0	4.00	19	-13	169	+ 26.0
1902.....	5.4	+ 1.4	1.96	37	+ 5	25	+ 7.0
1903.....	3.8	- 0.2	0.04	31	- 1	1	+ 0.2
1904.....	4.2	+ 0.2	0.04	32	+ 0	0
1905.....	5.0	+ 1.0	1.00	38	+ 6	36	+ 6.0
1906.....	3.0	- 1.0	1.00	37	+ 5	25	- 5.0
1907.....	5.6	+ 1.6	2.56	33	+ 1	1	+ 1.6
1908.....	3.3	- 0.7	0.49	30	- 2	4	+ 1.4
1909.....	4.7	+ 0.7	0.49	34	+ 4	16	+ 2.8
1910.....	4.6	+ 0.6	0.36	37	+ 5	25	+ 3.0
1911.....	2.6	- 1.4	1.96	32	+ 0	0
1912.....	4.0	+ 0.0	0	29	+ 7	49
1913.....	2.6	- 1.4	1.96	29	- 3	9	+ 4.2
1914.....	2.1	- 1.9	3.61	28	- 4	16	+ 7.6
1915.....	- 6.8	+ 2.8	7.84	31	- 1	1	- 2.8
		-16.5		-19	+109.9
		+13.5	48.56		+48	555	- 12.0
		- 3.0		- 1	+ 97.9
Base No. = 4.0.				32			
Mean = $4.0 + \frac{[x]}{28} = 3.89$				= $32 + \frac{[y]}{28} = 31.96$			

In a similar manner, using the symbols explained on page 563, the exact value of the sum of the products, xy , may be found from the computed sum by the following equation:

$$\Sigma xy = [xy] - \frac{[x][y]}{n}.$$

Accordingly, we have the following formulæ for computation of the coefficient of correlation:

Standard deviation (rainfall),

$$\sigma_x = \sqrt{\frac{[x^2] - \frac{[x]^2}{n}}{n}} = \sqrt{\frac{48.239}{28}} = 1.313 \text{ inches.}$$

Standard deviation (yield of corn),

$$\sigma_y = \sqrt{\frac{[y^2] - \frac{[y]^2}{n}}{n}} = \sqrt{\frac{554.96}{28}} = 4.452 \text{ bushels.}$$

$$\text{Coefficient of correlation, } r = \frac{[xy] - \frac{[x][y]}{n}}{n \sigma_x \sigma_y} = \frac{3.493}{5.846} = +0.598.$$

$$\text{Probable error of coefficient} = E_r = \pm 0.6745 \frac{1 - r^2}{\sqrt{n}} = \pm 0.12.$$

That is, $r = +0.69 \pm 0.12$, which magnitudes represent a fairly close order of correlation in problems of this character.

The equation of the straight line defined by the coefficient of correlation and expressing the direct influence of rainfall on the yield of corn as shown by the data analyzed is:

$$y = r \frac{\sigma_y}{\sigma_x} x = 2.027x.$$

The coefficient of x thus found, viz, 2.027, is identical with the one obtained in the direct least square computation, namely b , page 564. This agreement not only checks exactly all the arithmetical work, but shows the mathematical identity of the two methods of analysis.

The origin of coordinates for the line given by the equation $y = 2.027x$, is the point defined by the mean value of rainfall, viz, 3.89 inches, and the mean yield, viz, 31.96 bushels. The new axes are dotted in figure 10.

Variation of rainfall above and below the mean should, according to the indications of the data analyzed, be accompanied by corresponding changes of the yield above and below the average yield and in the proportion of 2.03 bushels of corn for each inch of July rainfall.

SUMMARY AND CONCLUSION.

I am indebted to Mr. William G. Reed for an interesting memorandum on the origin and history of the correlation coefficient and a short bibliography of a number of publications dealing with correlation and the theory of statistics. A portion of this is given in the note at the end. Some titles have also been added to the bibliography.

An effort has been made in this paper to outline in a general way the essential principles of the methods of least squares and the theories of statistics and correlation, with reference to their application in the analyses and presentation of climatic data and their utilization in the solution of problems of agricultural meteorology. While a considerable knowledge of mathematics is essential to a complete mastery of all the methods, processes and relations, nevertheless an elementary knowledge and a little study are sufficient to enable any one to carry out the relatively simple routine and systematized calculations that are necessary to bring out all the facts. Examples of these computations have been shown with considerable and seemingly all necessary fullness.

It is hoped the presentation will awaken interest in these valuable agencies and lead to their far greater application in climatic and agricultural studies. Such applications seem to be full of promise, and meteorologists can not afford to neglect, reject, or discredit either the methods or the results of the kind of studies herein considered.

It may seem at first thought that data is inadequate. A closer study indicates that lack of data is not necessarily a serious limitation. A record at a single station may be inadequate, but the meteorologist now has available an enormous mass of statistics which, properly grouped and combined, leaves little to be desired in fixing the general laws of variations and relations.

LITERATURE ON CORRELATION.

"The fundamental theorems of correlation were for the first time and almost exhaustively discussed by A. Bravais¹⁸ more than half a century ago. He deals completely with the correlation of two and three variables. Forty years later Mr. J. D. Hamilton Dickson¹⁹ dealt with a special problem proposed to him by Mr. Galton, and reached on a somewhat narrow basis some of Bravais' results for correlation of two variables. Mr. Galton at the same time introduced an improved notation which may be summed up in the 'Galton function' or coefficient of correlation. This indeed appears in Bravais' work, but a single symbol is not used for it. In 1892 Prof. Edgeworth, also unconscious of Bravais' memoir, dealt in a paper on 'Correlated Averages' with correlation for three variables.²⁰ He obtained results identical with Bravais', although expressed in terms of 'Galton's functions'."²¹

The following publications contain complete statements of the later developments and bibliographies are given where it is so indicated.

BIBLIOGRAPHY.

- Alvord, John W. & Burdick, [Charles S.]** Report to the Franklin County [Ohio] Conservancy District on flood relief for the Scioto valley. Alvord & Burdick, chief engineers. [Columbus, O., 1916.] [xii], 279 p. 65 figs. 8°. (front cover: State of Ohio. The Franklin County Conservancy District . . .) [See pp. 76-85 for "probable frequency of great floods"].
- Bravais, A.** Analyse mathématique sur les probabilités des erreurs de situation d'un point. Acad. des sci., Mémoires présentés par divers savants. Paris, 1846 (2) 9: 255-332.
- Brown, W.** The essentials of mental measurement. Cambridge, Univ. Press, 1911. — p. —°.
- Comstock, George C.** An elementary treatise upon the method of least squares with numerical examples of its application. Ginn, Boston, 1889°. 68 p. 8°.
- Davenport, C. B.** Statistical methods with special reference to biological variation. 3d rev. ed. Wiley, New York, 1914. viii, 225 p. 12°. [Bibliography on pp. 85-104.]
- Edgeworth, F. Y.** On correlated averages. Phil. mag., London, 1892 (5) 34: 190-204.
- Elderton, W. Palin.** Frequency-curves and correlation. C. & E. Layton, London. [1906]. xiii, 172p. 8°. (Published for the Institute of Actuaries.)
- Elderton, W. Palin, & Ethel M.** Primer of statistics. A. & C. Black, London, 1910. [See p. 55-72.]
- Galton, F.** Family likeness in stature. Proc., Roy. soc., London, 1886, 40: 63-73 (appendix).

- Great Britain. Meteorological Office.** The Computer's handbook. Section V.—Computations related to the theory of probabilities.
1. Errors of observation and variations due to accidental causes with an application to errors of means and normals. By R. Corless.
 2. Practical application of statistical methods to meteorology. By W. H. Dines.
- London, 1915. pp. VI-V52. 8°. (M. O., 223. Sect. V.)
- Hinrichs, Gustav Detlef.** Rainfall laws, deduced from twenty years of observations. Washington, 1893. 94 p. 16 fig. 8°. (U. S. Weather Bureau.)
- Hooker, R. H.** An elementary explanation of correlation; illustrated by rainfall and the depth of water in a well. Quart. jour., Roy. met'l. soc., London, 1908, 34: 277-291.
- Correlation of successive observations. Jour., Roy. statistical soc., London, —, 68: 676-703.
- Horton, Robert E.** Supplemental note on frequency of recurrence of Hudson River floods. In "The floods of 1913 . . . by Alfred J. Henry." Washington, 1913. 4°. (U. S. Weather Bureau bull. Z: W. B. no. 284.) pp. 109-112.
- Jacobs.** On crop yields, weather. (Indian meteorological department.)
- King, W. I.** Elements of statistical method. New York, 1912. pp. 197-215.
- Köppen, Wladimir.** Häufigkeit bestimmter Temperaturen in Berlin, verglichen mit frühen und heiteren Klimaten. Meteorol. Ztschr., Berlin, Juni, 1888, 5: 230-234.
- Merriman, Mansfield.** Method of least squares. New York, 1915. 8th ed. viii, 230 p. 8°.
- Pearson, Karl.** Contributions to the mathematical theory of evolution. London, Royal Society, Philosophical Transactions, Series A, as follows:
- (1) On the dissection of frequency curves, 1894, A185: 71-110.
 - (2) Skew variations in homogeneous material, 1895, A186: 343-414.
 - (3) Regression, heredity, and panmixia, 1896, A187: 253-318.
 - (4) On the probable errors of frequency constants and on the influence of random selection on variation and correlation, 1898, A191: 229-311.
 - (5) On the reconstruction of the stature of prehistoric races, 1898, A192: 169-244.
 - (6) Genetic (reproductive) selection; inheritance of fertility in man and of fecundity in thoroughbred race horses, 1899, A192: 257-330.
 - (7) On the correlation of characters not quantitatively measurable, 1900, A195: 1-47.
 - (8) On the inheritance of characters not quantitatively measurable, 1900, A195: 75-150.
 - (9) On the principle of homotypy and its relation to heredity, to the variability of the individual, and to that of the race, 1901, A197: 285-379.
 - (10) Supplement to a memoir on skew variation, 1901, A197: 443-459.
 - (11) On the influence of natural selection on the variability and correlation of organs, 1902, A200: 1-66.
 - (12) On a generalized theory of alternative inheritance, with special reference to Mendel's Laws, 1904, A203: 53-86.
- In London, Drapers' Company Research Memoirs, Biometric Series:
- (13) On the theory of contingency and its relation to association and normal correlation. Memoir 1.
 - (14) On the general theory of skew correlation and non-linear regression. Memoir 2.
 - (15) On the mathematical theory of random migration. Memoir 3, 1906.
 - (16) On further methods of determining correlation. Memoir 4, 1907.
 - (17) (Not published.)
 - (18) On a novel method of regarding the association of two variates classed solely in alternate categories. Memoir 7, 1912.
- Pearson, Karl.** Variation of the egg of the sparrow (*Passer domesticus*). Biometrika, Cambridge, January 1902, 1: 256-265.
- On the systematic fitting of curves to observations and measurements. Biometrika, Cambridge, April 1902, 1: 265-303; November 1902, 2: 1-23.
- Persons, W. W.** The correlation of economic statistics. Quart. publ., Amer. statist. assoc., Boston, 1910: 287-322.
- Reed, William Gardner & Tolley, Howard R.** Weather as a business risk in farming. Geogr. rev., New York, July 1916, 2: 48-53. illust. Abstract in MONTHLY WEATHER REVIEW, June 1916, 44: 354. 3 figs.
- Sheppard, W. F.** New tables of the probability integral. Biometrika, Cambridge, February 1903, 2: 174-190.
- Spillman, W. J., Tolley, H. R., & Reed, W. G.** The Average-interval curve and its application to meteorological phenomena. MONTHLY WEATHER REVIEW, April 1915, 44: 197-200. 2 figs.

¹⁸ Analyse mathématique sur les probabilités des erreurs de situation d'un point. Paris, Académie des Sciences, Mémoires présentés par divers savants. Series 2, v. 9, 1846, pp. 255-332.

¹⁹ Galton, F. Family likeness in stature. Proc. Royal Society, London, 1886, v. 40, Appendix, p. 63-73.

²⁰ Philosophical Magazine, London, Series 5, v. 34, 1892, p. 190-204.

²¹ Pearson, Karl. Philosophical Transactions, Royal Society, London, Series A, v. 187, 1896, p. 261.

- Sprung, A.** Ueber die Häufigkeit beobachteter Luft-Temperaturen in ihrer Beziehung zum Mittelwerthe derselben. *Meteorol. Ztschr.*, Berlin, April 1888, 5: 141-145. 1 fig.
- Stevens, James S.** *Theory of measurements.* Van Nostrand Co., New York, 1915. — p. —°.
- Walker, Gilbert T.** Correlation in seasonal variations of weather, I-VI. Simla, 1909-15. 1°. (Indian met'l. dept. Memoirs.)
- I. Correlation in seasonal variation of climate. v. 20, pt. 6, 1909, pp. 122-
- II. (A) On the probable error of a coefficient of correlation with a group of factors. v. 21, pt. 2, 1910, pp. 22-26.
(B) Some applications of statistical methods to seasonal forecasting. v. 21, pt. 2, 1910, pp. 26-45.
- III. On the criterion for the reality of relationships or periodicities. v. 21, pt. 9, 1914, pp. 13-16.
- IV. Sunspots and rainfall. v. 21, pt. 10, 1915, pp. 17-60.
- V. Sunspots and temperature. v. 21, pt. 11, 1915, p. 61-90.
- VI. Sunspots and pressure. v. 21, pt. 12, 1915, pp. 91-118.
- Weld, Leroy D.** *The theory of errors and least squares.* MacM. Co., New York, 1916. xii, 190 p. 12°.
- Yule, G. Udny.** *Introduction to the theory of statistics.* London, 1912. xiii, 376 p. 12°. [Bibliog. pp. 188, 208, 225, and 252.]

INJURY TO VEGETATION RESULTING FROM CLIMATIC CONDITIONS.¹

By GEORGE EDWARD STONE, Ph. D., Professor of Botany.

[Address: Massachusetts Agricultural College, Amherst, Mass.]

Nearly every winter furnishes conditions which are responsible for more or less injury to vegetation of both native and exotic species. During the past decade a vast amount of damage due to extreme conditions has resulted to vegetation, especially in the northeastern States. There has probably been no period within the memory of living men, or for that matter within the period of exact meteorological records, when damage to vegetation in America has been more extensive than during the past 12 years or since the winter of 1904. Every meteorological factor has its specific influence on vegetation, but since some of these influences are so intimately related to certain types of injury we will [sic] deal only with those concerned in the so-called winter injury. The principal meteorological factors associated with winter-killing and allied phenomena are temperature, soil and air moisture, wind, and light.

Either high or low temperatures or too much or too little soil moisture are conducive to abnormal conditions in plants; also the amount and intensity of light and the movements of the air form important factors in respect to this. Both winds and sunlight have a marked effect on transpiration, even sunlight alone greatly accelerating this process. Therefore, for a correct understanding of the cause underlying injury to vegetation from climatic conditions, it is essential to have some conception of the relative importance of meteorological agencies on plant development and the rôle which they play in regard to susceptibility to various troubles.

Some of the conditions which underlie winter-killing are as follows:

Severe and prolonged cold, causing frost to penetrate to a great depth.

Sudden and severe cold following a prolonged warm spell in the Fall, in which case the wood tissue may be tender and immature.

All conditions which favor a soft growth and immaturity of wood. Various causes may be responsible for this, such as growth in a low, moist soil, too heavy manuring or fertilization, or absence of sufficient sunlight.

General low vitality, caused by insect pests and fungous diseases and by lack of moisture in the soil.

Insufficient soil covering, such as lack of organic matter, light mulching and thin snow covering in winter.

Location in unusually windy and exposed places, etc.

A summer drought followed by copious rains during the Fall is often responsible for the production of immature tissue susceptible to cold.

Plants growing in the drainage of cesspools are likely to be affected by cold owing to the production of unripened wood.

Many of our introduced species are quite tender and are likely to be affected more or less every winter by severe cold. The buds of peach trees are generally affected by cold in the northern States and such plants as the privet, Japanese maples, etc., are affected by ordinary cold. On the other hand, plants that are native further north, such as the Labrador tea, frequently suffer some winter injury in our latitude when grown out of their natural environments. Swamp species transplanted to relatively dry soil suffer more from drought and low winter temperatures than those grown in their normal habitat. Many native plants are winter-killed badly when on the north side of buildings where light is insufficient, because in such situations the wood fails to mature properly. On the other hand, some southern species of plants, such as the magnolias, are more hardy in the north than are some of our native species. Indeed the reason why the magnolias do not grow more abundantly in the north is apparently not connected with temperature requirements.

Some injury to vegetation is generally caused by snow and ice, and this aside from that which occurs from the overloading of branches. The leaves of the lower branches of various conifers are often killed when buried in snow banks and the leaves of arbutus are commonly sun-scorched from exposure to winter snows and ice.

The injuries resulting to vegetation induced by meteorological conditions can be conveniently placed under two different categories, namely, injury to the root system and injury to the aerial portion of the plant, to limbs, branches, and leaves. Injuries which occur to a plant above the surface of the ground and which are associated with meteorological agencies are "frost cracks," "sun-scald," "sun-scorch," and "bronzing."

FROST-CRACKS.

"Frost-cracks" are formed in winter and are due to extreme changes in temperature within the tissues and occur on those portions of the tree where the maximum amount of heat is developed, namely, on the southwest side of the tree. Since the maximum amount of heat derived from sunshine is received generally between 2 and 3 p. m., that portion of a tree-trunk coinciding with the direct rays of the sun at this period is the one most likely to be affected on a day of uniform clearness. Moreover, the location of frost-cracks on a tree coincides with that area giving the minimum electrical resistance, and since the electrical resistance of a tree is proportional to the temperatures of the tissue comprising the same—the lower electrical resistance corresponding with the higher temperatures—that portion of the tree showing the least electrical resistance is most susceptible to frost-crack.

The opening and closing of frost-cracks are very responsive to changes in the meteorological conditions, they being influenced by variation in temperature, moisture, and in barometric conditions. They open more in winter than in summer, more under a dry than under a moist atmosphere, and more during high than during

¹ Reprinted from *Jour., New York botan. garden*, Oct. 1916, No. 202, 171:173-179.